高等学校教材

数理统计学讲义

第2版

北京大学

陈家鼎　孙山泽
李东风　刘力平　编著

高等教育出版社

内容提要

本版对初版的许多章节进行了改写与扩充，增添了许多重要的内容和实际应用例子，在叙述方法和内容编排上注意重点与非重点、基本内容与进一步内容的界限，层次分明，便于教学。全书共8章，包括绪论、估计、假设检验、回归分析与线性模型、试验设计与方差分析、序贯分析初步、统计决策与贝叶斯统计大意、抽样调查概述等，内容较丰富、重点突出、文字确切易懂。本书可作为高等学校数学专业、统计学专业及有关专业数理统计课程的教材或参考书。

第 2 版前言

根据我们在北京大学使用本教材(初版)的教学实践经验,并吸取了国内外近10年来出版的多部数理统计优秀教材(见参考书目)的一些好的讲法,本版对初版的许多章节进行了改写与扩充,以便在内容上和编排上更好地适应高等学校相关专业"数理统计"课程的教学需要及数理统计的应用需要。

本版的特点可概述如下。

(一) 增添了许多重要内容。例如

(1) 对寻找置信区间和置信限的方法进行了系统论述;

(2) 对"假设检验"中的 p 值方法进行了全面论述;

(3) 介绍了有关正态总体的 Behrens - Fisher 问题的解决方法;

(4) 介绍了比率的检验方法,特别是 Fisher 精确检验法;

(5) 对逻辑斯谛回归作了初步介绍;

(6) 对考虑交互作用的正交设计作了初步介绍;等等。

(二) 增加了许多实际应用例子。主要是增加了日常生活、社会调查、商务管理、医学试验等方面的例子。例如在讲比率的两样本检验时,介绍了某公安局两个刑侦组破案能力的比较;在讲多元回归时介绍了生理节律模型;在讲正交设计时,介绍了拖拉机噪声试验;等等。

(三) 在叙述方法和内容编排上注意重点与非重点、基本内容与进一步内容的界限,做到层次分明,便于教学。前四章是本书的主体部分。某些内容(包括新添加的一些内容)小字排印或标题上打上 * 号,以示这些内容不是最基本的,不要求学生掌握。"回归分析"的讲法有较大改变,先对一元线性回归作系统论述,然后再

讲线性模型和多元回归。考虑到有些读者不熟悉线性空间的投影理论，本版以附注的形式详细介绍了不用“投影理论”的处理方法。例如，用矩阵的基本运算直接证明了最小二乘估计的存在定理及高斯-马尔可夫定理。

本版仍坚持初版前言中叙述的原则：大学教材的内容应比教学大纲规定的多一些，更比课堂实际讲授的多一些。这样做有利于学生课外阅读，使有余力的学生选学更多的东西。

编者力图与时俱进，写出反映时代精神的合适教材。但限于水平，书中的缺点、谬误一定不少，欢迎读者和专家批评指出。

陈家鼎　孙山泽　李东风　刘力平

2003年12月于北京大学数学学院

初版前言

由于数理统计学在很多科学研究、工程技术和经济管理的领域里有广泛应用，许多高等院校都开设数理统计课。本书是根据笔者在北京大学概率统计专业的教学实践经验，在先前编写的同名讲义（北京大学油印本）的基础上进行较大的改写与扩充而成的。在编写过程中，我们注意了下列各点：

1）力图对数理统计学的基础部分进行比较准确而全面的系统论述，注意介绍统计思想和统计方法，特别重视数理统计学与社会实践的联系。

2）本书是入门读物，只要求读者学过微积分、线性代数和初等概率论。为了叙述数理统计学的现代发展，有少量内容涉及实变函数论，个别地方涉及测度论。这些涉及较深数学的内容不是本书的基本部分，不要求读者掌握。

3）我们认为，大学教材的内容应比教学大纲规定的多一些，更应比课堂实际讲授的多一些。这样做有利于学生课外阅读，使有余力的学生选学更多的东西。本书前四章是主体部分，要求学生掌握；后四章是介绍数理统计学的一些重要分支，教学中可根据教学时数选学部分内容，只要求学生有初步了解。我们把前四章中数学推理过于抽象或冗长的部分用楷体字排印，不要求学生掌握。

本书初稿完成后，承蒙中山大学邓永录教授、邓集贤教授、杨维权教授、华南师范大学汤尚勇教授、华东师范大学吕乃刚副教授进行了细心的审阅，他们提出了很多宝贵的意见。我们在修改和定稿时充分考虑了这些意见，在此向他们表示衷心感谢。高等教育出版社高尚华同志对本书的编写、审查与出版始终给予很大帮

助，在此一并致谢。

由于我们水平有限，本书一定有不少缺点和错误，欢迎各地专家和读者指正。

陈家鼎　孙山泽　李东风

1992年4月于北京大学概率统计系

本书常用的记号

R　　全体实数组成的集合，即 $R=(-\infty,\infty)$。

R^n　　n 维欧氏空间。

$\|a\|$　　向量 a 的模。当 $a=(a_1,\cdots,a_n)$ 或 $a=(a_1,\cdots,a_n)'$ 时，$\|a\|=\left(\sum_{i=1}^{n}a_i^2\right)^{\frac{1}{2}}$。

$I_E(x)$　　集合 E 的示性函数，即 $x\in E$ 时 $I_E(x)=1$，$x\overline{\in} E$ 时 $I_E(x)=0$。

$\stackrel{\mathrm{d}}{=}$　　按定义相等. $a\stackrel{\mathrm{d}}{=}b$ 表示用 b 来定义 a。

EX　　随机变量 X 的数学期望(均值)。

$\mathrm{Var}X$　　随机变量 X 的方差。

A'　　矩阵 A 的转置。

目 录

第一章 绪 论

§1 数理统计学的研究对象

数理统计学是数学的一个重要分支．它研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议．若在这句文字中去掉"带有随机性的"这几个字，那就是统计学的研究范围．统计学也就是数据科学．由于实践中人们收集到的数据大都受到随机性的影响，因而数理统计学是统计学的主要组成部分．在许多西方国家（如美国），这两个名词的含义基本相同，许多大学里统计系就是研究数理统计学的，使用"数理统计学"一词时只是强调统计学中用到很多现代数学．在我国由于历史的原因，这两个名词在实际使用时有较大差别．笔者相信，在若干年后，这两个名词所指的范围将相差不大．

数理统计学在工农业生产、工程技术、自然科学、经济学、社会学等领域都有很广泛的应用，大量实际问题属于数理统计学的研究范围．举几个例子：

（1）如何估计产品的寿命？这是工业品质量管理中极重要的问题．寿命试验是破坏性试验，只能抽取少量产品做试验．为了评价某批电子设备的使用寿命，随机抽取了 18 台做试验，测得寿命数据如下（单位：h）：17，29，50，68，100，130，140，270，280，340，410，450，520，620，190，210，800，1100．

问：整批电子设备中，寿命超过 200 h 的占多大比例？

（从第二章知，这个比例大约是 $e^{-\frac{200}{318}}=0.533$．）

有些寿命问题更为复杂，如导弹的贮存寿命如何确定？设某部门生产同一型号的一批导弹，制造后进行贮存，以备将来之用.导弹贮存太久功能就失效，变成废品.怎样合理地确定贮存期？这是很重要的实际问题.若随机抽取 n 件进行使用试验.设第 i 件已贮存 t_i 年（$i=1,2,\cdots,n$），试验结果无非是：该导弹的功能有效或无效.在前者情况下，第 i 个导弹的贮存寿命大于 t_i，在后者情况下第 i 个导弹的贮存寿命小于 t_i，在两种情况下，精确的贮存寿命是不知道的.有了上述类型的数据，如何确定该型号导弹的贮存期？

（2）某针织厂进行技术革新.在针织品的漂白工艺过程中，要考虑温度对针织品断裂强力（主要质量指标）的影响.为了比较 70 ℃与 80 ℃的影响有无差别，在这两个温度下分别作了 8 次试验，测得数据如下：

70 ℃时的强力：20.5，18.8，19.8，20.9，21.5，19.5，21.0,21.2；

80 ℃时的强力：17.7，20.3，20.0，18.8，19.0，20.1，20.2,19.1.

究竟两种温度下的强力有没有差别？可以算出这个试验中 70 ℃下的平均强力是 20.4,80 ℃下的平均强力是 19.4，这 1 个单位的差别主要是由于温度不同造成的还是由于随机误差造成的（注意，在同一温度下的数据波动也不小）？鉴别法是什么？

从第三章知道，这 1 个单位的差别主要是由于温度不同造成的，70 ℃时的强力确实大些.

（3）某公安局有两个刑侦组，在过去一年内第一组接手 25 件人命案，结果侦破了 23 件；第二组接手了 35 件人命案，结果侦破了 30 件，问：两个组的侦破能力有无差别？

从第三章知道，没有理由认为两个组的侦破能力有差别.

（4）在使用海洋重力仪探测海底石油时，仪器上显示出的信号 $x(t)$ 乃是重力加速度信号 $s(t)$ 与海浪干扰信号 $n(t)$ 的叠加：

$$x(t) = s(t) + n(t) \qquad (0 \leqslant t \leqslant T)$$

干扰信号 $n(t)$ 的强度一般是有用信号 $s(t)$ 的十万倍. 问:如何从强干扰背景中把微弱信号 $s(t)$ 提取出来? 也就是如何设计一个滤波器使得输入是 $x(t)(0 \leqslant t \leqslant T)$ 时输出是 $s(t)(0 \leqslant t \leqslant T)$?

这个问题已于 20 世纪 70 年代被北京大学和北京地质仪器厂的同志合作解决了.

(5) 在低钴定膨胀合金材料研究中,如何确定 Ni、Co、Cu、Fe 的合适比例以保证合金材料的膨胀系数与陶瓷材料的膨胀系数很接近? 如何做到尽量节省 Co 的用量?

这个问题可用数理统计中的回归分析法处理,已于 20 世纪 70 年代被北京大学和冶金部等单位的同志合作解决了,参看第四章.

(6) 为了探讨吸烟与患慢性支气管炎是否有关联,调查了 339 人. 情况如下表:

人　数	患慢性支气管炎	未患慢性支气管炎	合　计
吸　烟	43	162	205
不吸烟	13	121	134
合　计	56	283	339

问:从这批数据能否断定患慢性支气管炎与吸烟有关?

利用列联表分析法知答案是肯定的. 参看第三章.

从上面几个例子看出,数理统计能够处理的问题是多种多样的. 哪儿有受随机性影响的数据,哪儿就有数理统计的问题. 为了研究随机现象(偶然现象)中的规律性,就需要收集数据,并在此基础上对数据进行分析(由表及里、由此及彼),以得出科学的结论,这正是数理统计学的任务. 当然,和整个数学一样,数理统计学只是各个学科各个业务领域的研究工具,解决实际问题时不能离开所论问题的专门知识.

数理统计学含有丰富的统计方法与系统的数学理论. 它与整

个数学大厦的其他部分有密切关系. 数理统计学中用到很多近代数学知识,主要是微积分、函数论、线性代数,也用到泛函分析、拓扑学、组合数学、抽象代数和微分几何的知识,但与数理统计学关系最密切的是概率论. 在很大程度上可以说:概率论是数理统计学的基础,数理统计学是概率论的一种应用. 但它们是两个并列的数学分支学科,没有从属关系.

数理统计学又是一门实践性很强的应用科学,它同大量的数据打交道. 处理数据是一件十分重要而又相当繁琐的工作,电子计算机是数理统计学的重要工具. 每位优秀的数理统计工作者应该不仅具有良好的数学基础,而且能够熟练地使用电子计算机.

§2 数理统计学的基本概念

数理统计学的最基本概念是:总体、样本、统计量、估计、假设检验、统计决策等.

我们常把被研究的对象的全体(包括有形的和潜在的)称作总体. 常用随机变量(随机向量、随机过程)X 来表示一个总体(或总体的特性值),例如用 X 表示一批电子设备的寿命,意思是:从中任意抽取一件电子设备,其寿命是不能预先确定的,可看成随机变量 X 的值. 本讲义是数理统计学的入门教材,主要讨论随机变量的情形. 至于总体是随机向量或随机过程的情形,以后有"多元统计分析"和"时间序列分析"等课程专门进行论述.

以下以随机变量为例介绍若干基本概念(对于随机向量、随机过程的情形可做相仿的讨论). 当总体用随机变量 X 描述时,X 的分布函数 $F(x)$ 也称为总体的分布函数. 要了解总体就是要了解其分布函数. 在数理统计学中,分布函数 $F(x)$ 是未知的,研究工作就是要对总体进行调查(或观测)并对调查来的数据进行科学处理,以便知道 $F(x)$ 是什么样的.

怎样进行调查? 逻辑上有两种方式:普查和抽查. 普查就是

对总体的每个个体都进行调查，这种方式有很大的局限性，当总体包含的个体数相当大时，由于调查工作量太大而难以采用；当总体包含的个体有无限多或对个体的调查具有破坏性时，根本不能采用. 另外一种方式就是抽查，即从总体中抽取若干个有代表性的个体 $X_1,X_2,\cdots,X_n$（这里 X_i 是第 i 次抽取时得到的个体的观测值），$(X_1,\cdots,X_n)$叫做总体的样本，n 叫做样本量或样本大小. 怎样抽查呢？数理统计学中采取“随机抽样法”. 即样本中包含哪些个体不是事先确定的，总体中的每一个体均有机会进入样本. 在数理统计学中使用得最多的是“简单随机抽样法”. 这个名词有两种不同的含义①，其中最常用的一种含义是：它是这样的一种抽取方法，使得 $X_1,\cdots,X_n$ 作为 n 个随机变量是相互独立的，而且与总体有相同的概率分布. 这种样本叫做“简单随机样本”.

怎样才能得到“简单随机样本”呢？有两种基本方法. ①“有放回地逐次随机抽取法”. 总体中的每个个体都有同样的机会被抽入样本，且每次抽出的个体，在记下其值后，还要放回到总体中去，以保证在下次抽取时每个个体仍有与第一次抽取时相等的机会被抽入样本. 随机性表现在：样本中包含哪些个体，是出自机会，而不是在抽样前预定的. ② 对总体 X 进行多次独立的重复观测，这时观测到的值可以看成是总体的所有可能值（无形地存在着）的一部分. 例如用仪器对某一物体的长度进行精密测量，我们把测量结果看成随机变量（总体可想像为一切可能值的集合，例如$(0,\infty)$或更大的集合），把 n 次重复测量的结果记为 $X_1,\cdots,X_n$，则得到简单随机样本.

我们强调指出，在实施抽样之后得到 n 个实数 $x_1,\cdots,x_n$. 它

---

① 在总体只含有限个个体的情形（如有 N 个个体），“简单随机抽样法”往往指：从总体中随机地抽取 n 个. 这里“随机地”含义是：从 N 个个体中任意抽取 n 个，共有 C_N^n个可能的结果，这些结果有相等的概率，都是$(C_N^n)^{-1}$. 这是无放回的抽取法，得到的样本$(X_1,\cdots,X_n)$不是下文定义的“简单随机样本”，我们称之为“单纯随机样本”，以示区别. 对这种样本，在第八章中进行讨论.

们分别是 $X_1,\cdots,X_n$ 的观测值，通常把 $x_1,\cdots,x_n$ 叫做样本值，也简称为样本. 为了使用方便，同一字母有时代表随机变量，有时代表随机变量取的值. 只要看清上下文，是不会引起混乱的. 望初学者加以注意.

本讲义主要研究"简单随机样本"，以后简称随机样本或样本.

数理统计学就是采取随机抽样法获取样本，并从样本推断总体的性质. 统计推断的类别很多，主要有两类，一类是估计，另一类是假设检验. 下面是这两类问题的一些提法.

估计. 设总体 X 的分布函数 $F(x)$ 是未知的，如何根据样本值 $x_1,\cdots,x_n$ 给出 $F(x)$（一切 x）的估计值？

设 X 的分布函数为 $F(x,\theta)$，$\theta=(\theta_1,\cdots,\theta_m)$ 是未知向量，如何根据样本值估计出 θ？

更一般地，设 $g(F)$ 是分布函数 $F(x)$ 的泛函，如何根据样本值估计出 $g(F)$？

设 X 与 Y 是两个相互独立的随机变量，如何根据 (X,Y) 的样本值 $(x_1,y_1),\cdots,(x_n,y_n)$ 估计概率 $P(X>Y)$？

若 $X_1,X_2,\cdots$ 是一个随机序列，如何根据 $X_1,\cdots,X_n$ 的观测值 $x_1,x_2,\cdots,x_n$ 预测 X_{n+m}（m 是正整数）的值？

估计问题是多种多样的. 我们把预测问题看成特殊的估计问题. 估计问题的特点是找出合适的估计量 $\varphi(X_1,\cdots,X_n)$，作为被估计量的近似值.

假设检验. 设总体 X 的分布函数 $F(x)$ 是未知的，$\mathscr{F}_0$ 是一些分布函数组成的集合. 问：$F(x)$ 是否属于 $\mathscr{F}_0$？ 或者问看法（即假设）

$$F(x)\in\mathscr{F}_0 \tag{2.1}$$

是否可以接受？要根据 X 的样本值 $x_1,\cdots,x_n$ 给予明确的回答. 换句话说，在 $\mathbf{R}^n$ 中要确定一个集合 $W\subset\mathbf{R}^n$，当 $(x_1,\cdots,x_n)\in W$ 时拒绝看法(2.1)；否则，接受看法(2.1)，这个 W 叫做否定域. 怎样选择合适的 W？

设 X 与 Y 是两个随机变量，它们的分布函数分别是 $F(x)$ 与 $G(x)$. 问：如何根据(X,Y)的样本值$(x_1,y_1),\cdots,(x_n,y_n)$判断 X 与 Y 是否相互独立？又问：如何判断是否成立 $F(x)\equiv G(x)$(一切 x)？

假设检验问题的提法也是多种多样的，在抽样验收、科学试验、质量管理、民意测验等很多领域中要用到.

通常称样本 $X_1,\cdots,X_n$ 的函数 $g(X_1,\cdots,X_n)$为统计量，这里 g 是完全已知的函数[①]，一旦样本 $X_1,\cdots,X_n$ 的值取定，统计量的值就可以算出来了. 在估计问题里，估计量就是统计量，在假设检验问题里，否定域也常由统计量表达出来. 把样本看成随机变量时，统计量也是随机变量，它的分布有时叫做抽样分布.

估计问题、假设检验问题可看成是更一般的统计决策问题的特殊情形. 在第七章中对统计决策问题有简单介绍.

§3 数理统计学发展简史

数理统计学的发展大致可分三个时期来叙述.

20 世纪以前，这是数理统计学的萌芽时期. 在这漫长的时期里，描述性统计占据主导地位. 描述性统计就是收集大量的数据，并进行一些简单的运算(如求和，求平均值，求百分比等等)或用图表、表格把它们表示出来. 中国古代就有钱粮户口的统计，西方国家也多次进行人口统计，早期这些统计工作都与国家实施统治有关. 统计学的英文词 statistics 源出于拉丁文，系由 status(状态、国家)和 statista(政治家)衍化而来. 这时期也出现了一些现在仍很常用的统计方法，如直方图法. 但最重要的、超出描述性统计范围的成就是高斯和勒让德关于最小二乘法的工作. 在统计思想上

---

① 为了进行严密的数学研究，通常假定 g 是 Borel 可测函数. 学过实变函数论的读者知道，Borel 可测函数是极为广泛的一类函数，我们在实践中遇到的函数一般都是 Borel 可测函数.

的重大进展是:数据是来自服从一定概率分布的总体,而统计学就是用数据去推断这个分布的未知方面. 这个观点强调了推断的地位,使统计学摆脱了单纯描述的性质. 由于高斯等人在误差方面的研究工作,正态分布(又叫高斯分布)的性质和重要性受到广泛重视. 19 世纪末皮尔逊(K. Pearson, 1857—1936)引进了一个以他的名字命名的分布族,它包含了正态分布及现在已知的一些重要的非正态分布,扩大了人们的眼界. 皮尔逊还提出了一个估计方法——矩估计法,用来估计他所引进的分布族中的参数.

另外,德国的大地测量学者赫尔梅特(F. Helmert)1876 年在研究正态总体的样本方差时,发现了十分重要的χ^2分布. 高尔顿(F. Galton)在生物学研究中提出了回归分析方法. 这些都是数理统计学发展史中的重要事件.

20 世纪初到第二次世界大战结束,这是数理统计学蓬勃发展达到成熟的时期. 许多重要的基本观点和方法,以及数理统计学的主要分支学科,都是在这个时期建立和发展起来的. 在其发展中,以费希尔(R. A. Fisher, 1890—1962)为代表的英国学派起了主导的作用.

K.皮尔逊在 1900 年提出了检验拟合优度的χ^2统计量,并证明其极限分布(在原假设成立时)是χ^2分布. 这个结果是大样本统计的先驱性工作. 20 世纪 20 年代费希尔又作了重要发展.

紧接着的一项重要进展,是皮尔逊的学生,英国医生戈塞特(W. S. Gosset,笔名"学生")1908 年导出了 t 统计量的精确分布——t 分布,开了小样本理论的先河.

比皮尔逊略晚的费希尔对现代数理统计的形成和发展作出了最大的贡献,他是一些重要统计分支和方法的开创者. 最重要的成就有:系统地发展了正态总体下种种统计量的抽样分布理论,这标志着相关、回归分析和多元分析等分支的建立;建立了以最大似然估计为中心的点估计理论;创立了实验设计,并发展了相应的数

据分析方法——方差分析.

在数理统计学的另一个主要分支——假设检验的发展中，费希尔也起过重要作用，但假设检验理论的系统化和深入研究，则应归功于奈曼（J. Neyman，1894—1981）和小皮尔逊（E. S. Pearson，1895—1980）. 他们在1928—1938年期间发表了一系列文章，建立了假设检验的严格数学理论. 奈曼对数理统计作出的另一项很重要的贡献，是在1934—1937年间建立了置信区间的数学理论. 它基于概率的频率解释，并与奈曼-皮尔逊的假设检验理论有密切关系.

多元统计分析是数理统计学中有重要应用价值的分支之一. 1928年以前，费希尔在多元正态总体的统计分析中作过基础性的工作. 1928年威沙特（U. W. Wishart）导出了著名的"威沙特分布". 此后，这门统计分支发展很快. 我国著名学者许宝騄（1910—1970）对这一领域以及线性模型的统计推断理论，作出了奠基性的工作. 时间序列分析是数理统计学中有重要应用价值的又一分支. 尤尔（G. U. Yule）在1925—1930年间引进了自回归和序列相关等重要概念，奠定了这个分支现代发展的基础. 维纳（N. Wiener）在第二次世界大战期间为研究大炮射击问题而提出的时间序列分析方法（外推、内插和光滑化），不但对炮兵而且对整个工程界都有重要意义. 瓦尔德（A. Wald，1902—1950）在第二次世界大战期间提出了序贯概率比检验法，不仅在实用上有重要意义（例如贵重产品的抽样检查与验收），也为战后序贯分析的发展奠定了基础.

如上所述，在这不到半个世纪的时期里，数理统计学得到了丰富的多方面的发展，确立了这门学科在人类文化史上的地位. 克拉默（H. Cramer）在1946年写的名著"统计学的数学方法"对这一时期的主要成就作了扼要的总结，是此后20年间数理统计学的主要教科书.

战后时期：在这一时期中，数理统计学在应用和理论两方面继

续获得很大发展. 在应用上,由于经济和军事技术的快速发展以及电子计算机的出现,使数理统计学的应用达到了前所未有的规模. 有些需要大量计算的统计方法,在战前限于条件而无法使用,这个障碍今日已不复存在. 统计方法已渗透进各种专门的学科领域,形成了许多边缘学科,如统计质量管理、生物统计、气象统计、地质统计、计量经济学、医学统计等.

战后的统计理论沿着纵深方向快速发展. 例如最大似然估计和非参数统计的大样本理论,在战前只有初步的结果,现已达到完善的地步. 瓦尔德在 1950 年创立了统计决策理论,它从人与大自然进行博弈的观点出发,企图把形形色色的统计问题纳入一个统一的模式之下,这种理论对战后数理统计许多分支的发展产生了很大影响,特别是参数估计这个分支.

随着概率论的高速发展,随机过程的统计逐步形成了内容丰富的重要分支. 其中非线性滤波理论占据了显著地位,它是 40 年代维纳-柯尔莫戈罗夫滤波理论和 60 年代卡尔曼滤波理论向非线性领域的扩展. 前苏联学者李普泽尔(R. S. Liptser)和希拉也夫(A. N. Shiryayev)在 1974 年写的专著"随机过程的统计"系统论述了这方面的理论.

在战后数理统计的发展中,一个引人注目的现象是贝叶斯学派的崛起. 这个学派的思想可溯源于 1763 年发表的著名论文"论有关机遇问题的求解". 论文的作者是英国牧师贝叶斯(T. Bayes, 1702—1761). 这篇文章基于归纳推理的思想,当时未引起多大反响,却在 180 年后重新引起人们的注意. 20 世纪 50 年代以来,这个学派的兴起不是没有原因的,传统的统计学发展趋于成熟并得到大量应用后,其固有的弱点开始显露并逐渐为人们所认识. 传统统计方法不能充分利用过去经验积累起来的知识,在小样本问题里表现出难以克服的局限性. 这一点在可靠性统计问题中特别突出. 而在贝叶斯学派那里,却有独到的方法,利用过去积累的经验,在应用中常得到意想不到的效益. 虽然如此,实行贝叶斯方法

仍存在很多困难,先验分布的客观性常引起非议,贝叶斯学派的观点还难以为广大统计工作者普遍接受,贝叶斯学派和传统学派的争论仍将长期存在,就目前情况而言,后者大体上仍处于支配地位.

战后数理统计的发展中又一显著特点是电子计算机的广泛使用. 这不仅使得过去难以计算的问题能够解决,而且促使能有效地利用现代计算机强大计算能力的统计学新理论问世. 自助法(bootstrap)和投影寻踪法(projection pursuit)就是统计学工作者颇感兴趣的两种方法. 1980 年以来,对它们的应用范围和理论基础,统计学家做了许多研究工作.

以上对数理统计学的历史作了简短的叙述. 读者如想知道更多的历史知识,请参看陈希孺著的《数理统计学简史》(湖南教育出版社,2002).

第二章 估 计

§1 参数估计的方法

估计问题是多种多样的，很难纳入一个统一的模式. 本章讨论的估计问题的类型是比较简单的，首先讨论总体参数的估计，然后讨论分布函数、密度函数的估计.

设随机变量 X 的分布函数是 $F(x,\theta)$，密度函数或概率函数是 $f(x,\theta)$（当 X 是连续随机变量时，$f(x,\theta)$是通常的概率分布密度；当 X 是离散随机变量时，$f(x,\theta)$是概率函数，即 X 取值等于 x 的概率），其中 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$是未知的参数（向量），$\theta$ 的一切可能的值的集合记为 Θ，这个 Θ 是 m 维欧氏空间 R^m 的已知非空子集. 设 $g(\theta)$是 θ 的实值函数（通常是连续函数）. 问：如何根据 X 的简单随机样本 $x_1,x_2,\cdots,x_n$ 估计 $g(\theta)$的值？

例如，设 $X\sim N(\mu,\sigma^2)$（正态分布），μ,σ^2 是未知参数，$\theta=(\mu,\sigma^2)\in\Theta=(-\infty,\infty)\times(0,\infty)$. 如何估计 $P(X>c)$？这是上述类型的问题. 实际上 $P(X>c)=1-\Phi\left(\frac{c-\mu}{\sigma}\right)$，这里 $\Phi(x)$是标准正态分布函数，即

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

估计 $g(\theta)$的关键是如何估计 $\theta_1,\theta_2,\cdots,\theta_m$，若它们的估计值分别是 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_m$，则 $g(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_m)$就可作为 $g(\theta)$的估计值.

如何获得 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$的估计呢？方法很多，最重要最常用的有两种，一是最大似然（估计）法，另一是矩（估计）法，分别

介绍如下.

(一) 最大似然法

这是著名英国统计学家 R. A. Fisher 在 1912 年提出的. 设 X 的密度函数(或概率函数)是 $f(x,\theta)$,则简单随机样本(X_1, X_2,…,X_n)的联合密度函数(或概率函数)是

$$L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n}{}^{①} f(x_i,\theta) \tag{1.1}$$

当(X_1,…,X_n)取定值(x_1,…,x_n)时,$L(x_1,\cdots,x_n;\theta)$是 θ 的函数,这个函数叫做样本的似然函数. 若 $\hat{\theta}=\varphi(x_1,\cdots,x_n)$满足

$$L(x_1,\cdots,x_n;\hat{\theta})=\sup_{\theta\in\Theta} L(x_1,\cdots,x_n;\theta)$$

则称 $\hat{\theta}$ 是 θ 的最大似然估计②,也称 $\hat{\theta}$ 的各分量是 θ 的相应分量的最大似然估计.

以下记 $L(\theta)=L(x_1,\cdots,x_n;\theta)$. 由于 $\ln x$ 是 x 的严格增函数,故 $L(\theta)$的最大值点 $\hat{\theta}$ 也是 $\ln L(\theta)$的最大值点. 怎样求得最大似然估计 $\hat{\theta}$ 呢? 当 Θ 是开集而且 $\ln L(\theta)$的偏导数(导数)存在时,从微积分知识知,最大似然估计 $\hat{\theta}$ 应满足方程组:

$$\frac{\partial \ln L(\theta)}{\partial \theta_1}=0,\cdots,\frac{\partial \ln L(\theta)}{\partial \theta_m}=0 \tag{1.2}$$

(1.2)称为似然方程组. 从这个方程组求出根后,再判断根是不是似然函数的最大值点. 要注意:当 $L(\theta)$不存在偏导数时,需要直接研究 $L(\theta)$,寻找最大值点.

最大似然估计的根据是什么? 采用最大似然估计合理吗? 我们来考虑一个简单的情况. 假定一个盒子里有白球、黑球共 3 个,但不知白球、黑球各有几个. 如果有放回地从盒子里抽 3 个球,发现第 1、第 3 个是黑的,第 2 个是白的. 问:如何估计盒中黑球所占

---

① 在不引起混淆时可简记为 $\prod_{1}^{n}$.

② 为了便于数学上处理,我们要求 $\hat{\theta}=\varphi(x_1,\cdots,x_n)$是 $x_1,\cdots,x_n$ 的 Borel 可测函数,这个要求在实际工作中一般都是满足的.

的比例 θ? 不难看出，θ 的可能值的集合 $\Theta=\left\{0,\frac{1}{3},\frac{2}{3},1\right\}$. 从盒中有放回地取球，第 i 次取出的结果记为

$$X_i=\begin{cases}1, & \text{取出的是黑球}\\ 0, & \text{取出的是白球}\end{cases}\quad (i=1,2,3)$$

则 (X_1,X_2,X_3) 的联合概率函数为

$$L(x_1,x_2,x_3;\theta)=\theta^{x_1+x_2+x_3}(1-\theta)^{3-(x_1+x_2+x_3)}$$

(这里 $x_i=0$ 或 $1,i=1,2,3$).

当 (X_1,X_2,X_3) 取定结果为"第 1、第 3 个是黑球，第 2 个是白球"，即 $x_1=x_3=1,x_2=0$ 时，

$$L(\theta)=\theta^2(1-\theta)$$

$L(\theta)$ 是盒中黑球所占比例为 θ 时，事件"第 1、第 3 次取出的是黑球，第 2 次取出的是白球"的概率. 对应 θ 的各种可能值，这个概率如下：

θ	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$L(\theta)$	0	$\frac{2}{27}$	$\frac{4}{27}$	0

既然"第 1、第 3 次取出的是黑球，第 2 次取出的是白球"是已经确实发生了的事件，故与这一事实最匹配的 θ 值应该是使其出现概率取最大值的 $\theta=\frac{2}{3}$ 了. 这就是 θ 的最大似然估计.

上述解释完全适用于任何离散型随机变量情形，我们也可推广到有通常密度函数的情形. 设观察到样本值 $X_1=x_1,\cdots,X_n=x_n$，实际上表明观察到事件 $A=\{X_1\in(x_1-\delta,x_1+\delta),\cdots,X_n\in(x_n-\delta,x_n+\delta)\}$ 发生(对一切正数 δ). 易知

$$\begin{aligned}P(A)&=\int_{x_1-\delta}^{x_1+\delta}\cdots\int_{x_n-\delta}^{x_n+\delta}\prod_1^n f(u_i,\theta)\mathrm{d}u_1\cdots\mathrm{d}u_n\\ &\approx L(x_1,\cdots,x_n;\theta)(2\delta)^n\end{aligned}$$

（当 δ 很小时）可见，若用最大似然估计 $\hat{\theta}$ 代替 θ 时 $P(A)$ 达到最大.

以上是从直观上理解最大似然估计的合理性. 这种概率最大原则是一种朴素而重要的统计思想，应用很广. 数学上可以证明，在一定条件下（这些条件在大多数实际工作中是满足的），只要样本量 n 足够大，最大似然估计 $\hat{\theta}_n$ 与真值 θ 可相差任意小，即对任意 $\varepsilon>0$ 有

$$\lim_{n\to\infty} P(\|\hat{\theta}_n-\theta\|<\varepsilon)=1 \tag{1.3}$$

这里 $\|\hat{\theta}_n-\theta\|$ 表示 $\hat{\theta}_n$ 与 θ 的距离（欧氏距离）. 比（1.3）更强的性质是

$$P(\lim_{n\to\infty}\hat{\theta}_n=\theta)=1 \tag{1.4}$$

若（1.3）成立，通常称 $\hat{\theta}_n$ 有相合性；若（1.4）成立，通常称 $\hat{\theta}_n$ 有强相合性. 这方面的一般理论需要用到较多的数学知识（如测度论）.

下面通过一些例子说明最大似然估计的求法与相合性.

（1）两点分布 （也称伯努利（Bernoulli）分布）

设随机变量 X 取值 0 或 1，且

$$p=P(X=1)=1-P(X=0)$$

于是 X 的概率函数是

$$P(X=x)=p^x(1-p)^{1-x}, \quad (x=0,1)$$

其中 p 未知，$p\in[0,1]$. 设样本是 $x_1,x_2,\cdots,x_n$，则似然函数是

$$L(x_1,\cdots,x_n;p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_1^n x_i}(1-p)^{n-\sum_1^n x_i}$$ ①

取对数并对 p 求导后知 $\hat{p}=\frac{1}{n}\sum_1^n x_i=\bar{x}$ 是唯一的最大值点，即它

---

① $\sum_1^n x_i$ 是 $\sum_{i=1}^n x_i$ 的简写.

是 p 的最大似然估计. 直观上,如果某随机事件发生的概率是 p,则在 n 次独立重复试验中,此事件发生的频率为 $\frac{1}{n}\sum_{1}^{n}x_i=\hat{p}$,正好是最大似然估计. 由强大数律知 $\hat{p}$ 有强相合性.

(2) 指数分布

密度函数是

$$f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\lambda\in(0,\infty)$. 样本 $x_1,x_2,\cdots,x_n$ 的似然函数为

$$L(\lambda)=\lambda^n\prod_{1}^{n}e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{1}^{n}x_i}$$

于是

$$\frac{d\ln L(\lambda)}{d\lambda}=\frac{n}{\lambda}-\sum_{1}^{n}x_i$$

故似然方程的根 $\hat{\lambda}=n\Big/\sum_{1}^{n}x_i$. 不难看出 $\hat{\lambda}$ 是 λ 的最大似然估计.

注意总体的数学期望为 λ^{-1}. 依强大数律知 $\hat{\lambda}$ 是强相合的.

例 1.1 已知某种电子设备的使用寿命(从开始使用到失效为止)服从指数分布,密度函数 $f(x,\lambda)=\lambda e^{-\lambda x}(x>0,\lambda>0)$. 今随机抽取 18 台,测得寿命数据如下(单位:h):

17,29,50,68,100,130,140,190,210

270,280,340,410,450,520,620,800,1 100

问:如何估计 λ?

采用最大似然法. 利用公式 $\hat{\lambda}=\frac{n}{\sum_{1}^{n}x_i}$,现在 $n=18$, $\sum_{1}^{n}x_i=5\ 724$,故 $\hat{\lambda}=\frac{1}{318}$.

(3) 正态分布

密度函数是

$$f(x,\mu,\delta)=\frac{1}{\sqrt{2\pi\delta}}e^{-\frac{1}{2\delta}(x-\mu)^2}$$

其中 $\delta=\sigma^2\in(0,\infty)$，$\mu\in(-\infty,\infty)$. 样本(值) $x_1,\cdots,x_n$ 对应的似然函数为

$$L(x_1,\cdots,x_n;\mu,\delta)=\left(\frac{1}{\sqrt{2\pi\delta}}\right)^n e^{-\frac{1}{2\delta}\sum_1^n(x_i-\mu)^2}$$

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\delta-\frac{1}{2\delta}\sum_1^n(x_i-\mu)^2$$

似然方程组为

$$\frac{\partial\ln L}{\partial\mu}=\frac{1}{\delta}\sum_1^n(x_i-\mu)=0$$

$$\frac{\partial\ln L}{\partial\delta}=-\frac{n}{2\delta}+\frac{1}{2\delta^2}\sum_1^n(x_i-\mu)^2=0$$

其根是 $\hat{\mu}=\frac{1}{n}\sum_1^n x_i=\overline{x}$，$\hat{\delta}=\frac{1}{n}\sum_1^n(x_i-\overline{x})^2$. 可以验证似然函数 L 确实在 $(\hat{\mu},\hat{\delta})$ 处达最大值，所以 $\hat{\mu},\hat{\delta}$ 分别是 μ,δ 的最大似然估计.

由于总体的期望 $EX=\mu$，方差 $\mathrm{Var}(X)=EX^2-(EX)^2=\delta$，从强大数律知 $P(\lim\limits_n$① $\hat{\mu}=\mu)=1$，而

$$\hat{\delta}=\frac{1}{n}\sum_1^n x_i^2-\left(\frac{1}{n}\sum_1^n x_i\right)^2$$

再利用强大数律知

$$P(\lim_n\hat{\delta}=EX^2-(EX)^2)=1$$

故 $P(\lim\limits_n\hat{\delta}=\delta)=1$. 这就证明了 $\hat{\mu},\hat{\delta}$ 分别是 μ,δ 的强相合估计.

(4) 韦布尔(Weibull)分布

密度函数是

---

① $\lim\limits_n$ 是 $\lim\limits_{n\to\infty}$ 的简写.

$$f(x;m,\eta)=\begin{cases}\dfrac{m}{\eta^m}x^{m-1}\mathrm{e}^{-\left(\frac{x}{\eta}\right)^m}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $m>0,\eta>0$ 都是未知参数.

样本 $x_1,x_2,\cdots,x_n$ 对应的似然函数为

$$L(m,\eta)=\left(\frac{1}{\eta^m}\right)^n m^n\prod_1^n x_i^{m-1}\cdot \mathrm{e}^{-\frac{1}{\eta^m}\sum\limits_1^n x_i^m}$$

似然方程组为

$$\begin{cases}\dfrac{\partial \ln L}{\partial m}=-n\ln\eta+\dfrac{n}{m}+\sum\limits_1^n\ln x_i-\dfrac{1}{\eta^m}\sum\limits_1^n x_i^m\ln x_i\\ \qquad+\dfrac{1}{\eta^m}(\ln\eta)\sum\limits_1^n x_i^m=0 & (1.5)\\ \dfrac{\partial \ln L}{\partial \eta}=-\dfrac{nm}{\eta}+\dfrac{m}{\eta^{m+1}}\sum\limits_1^n x_i^m=0 & (1.6)\end{cases}$$

从(1.6)得

$$\eta=\left(\frac{1}{n}\sum_1^n x_i^m\right)^{\frac{1}{m}} \tag{1.7}$$

再代入(1.5)得

$$\frac{1}{m}+\frac{1}{n}\sum_1^n\ln x_i-\frac{1}{\sum\limits_1^n x_i^m}\sum_1^n x_i^m\ln x_i=0 \tag{1.8}$$

可以证明,当 $n\geqslant 2$ 且 $x_1,\cdots,x_n$ 不全相等时,方程(1.8)恰有一个根 $\hat{m}$. 再由(1.7)得

$$\hat{\eta}=\left(\frac{1}{n}\sum_1^n x_i^{\hat{m}}\right)^{\frac{1}{\hat{m}}}$$

这$(\hat{m},\hat{\eta})$便是韦布尔分布参数(m,η)的最大似然估计($(\hat{m},\hat{\eta})$不仅

是似然方程组的解，而且似然函数 L 在 $(\hat{m},\hat{\eta})$ 达最大值！请读者自己验证这一点）．与指数分布、正态分布的情形不同，这里的 $\hat{m}$ 没有明显的数学表达式，要找它就需解超越方程(1.8)．由于(1.8)式左端是 m 的严格减函数，利用计算机可以迅速地得到 $\hat{m}$ 的近似值.

数学上可以证明，$\hat{m},\hat{\eta}$ 分别是 m,η 的强相合估计.

(5) 均匀分布

密度函数是

$$f(x;a,b)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}$$

其中 $a<b$ 是两个未知参数.

样本 $x_1,\cdots,x_n$ 对应的似然函数为

$$L(x_1,\cdots,x_n;a,b)=\left(\frac{1}{b-a}\right)^n\prod_1^n I_{[a,b]}(x_i),$$

这里 $I_E(x)$ 表示集合 E 的示性函数($x\in E$ 时 $I_E(x)=1$；$x\overline{\in}E$ 时 $I_E(x)=0$)．不难看出，

$$L(x_1,\cdots,x_n;a,b)=\begin{cases}\left(\dfrac{1}{b-a}\right)^n, & \text{当 } a\leqslant\min\limits_{1\leqslant i\leqslant n}x_i \text{ 且}\\ & \quad b\geqslant\max\limits_{1\leqslant i\leqslant n}x_i\\ 0, & \text{其他}\end{cases}$$

故 a,b 的最大似然估计分别是 $\hat{a}=\min\limits_{1\leqslant i\leqslant n}x_i,\hat{b}=\max\limits_{1\leqslant i\leqslant n}x_i$.

数学上可以证明，$\hat{a},\hat{b}$ 分别是 a,b 的强相合估计．实际上，设与 $f(x;a,b)$ 相应的分布函数是 $F(x)$，则 $\hat{b}$ 的分布函数是 $[F(x)]^n$．设 $0<\varepsilon<b-a$，则 $P(|\hat{b}-b|\geqslant\varepsilon)=P(a\leqslant\hat{b}\leqslant b-\varepsilon)=[F(b-\varepsilon)]^n$

$= \left(\frac{b-a-\varepsilon}{b-a}\right)^n$，故 $\sum_{n=1}^{\infty} P(|\hat{b}-b| \geqslant \varepsilon) < \infty$. 于是不难推知①
$P(\lim_n \hat{b} = b) = 1$，即 $\hat{b}$ 是强相合的.

由于

$$\begin{aligned} P(\hat{a} < x) &= 1-[1-F(x)]^n, \\ P(|\hat{a}-a| \geqslant \varepsilon) &= P(a+\varepsilon \leqslant \hat{a} \leqslant b) \\ &= [1-F(a+\varepsilon)]^n \\ &= \left(1-\frac{\varepsilon}{b-a}\right)^n, \end{aligned}$$

故 $\sum_{n=1}^{\infty} P(|\hat{a}-a| \geqslant \varepsilon) < \infty$，于是 $P(\lim_n \hat{a} = a) = 1$，即 $\hat{a}$ 是 a 的强相合估计.

(二) 矩法

这是著名英国统计学家 K. Pearson 在 1894 年提出来的. 设随机变量 X 的密度函数(或概率函数)是 $f(x,\theta)$，其中 $\theta=(\theta_1,\cdots,\theta_m)$是未知参数. 若 X 的 k 阶原点矩 $V_k = EX^k$ 存在，则 V_k 也是 $\theta_1,\cdots,\theta_m$ 的函数，记为 $g_k(\theta_1,\cdots,\theta_m)(k=1,2,\cdots)$. 假定从方程组

$$\begin{aligned} g_1(\theta_1,\cdots,\theta_m) &= V_1 \\ g_2(\theta_1,\cdots,\theta_m) &= V_2 \\ &\cdots\cdots \\ g_m(\theta_1,\cdots,\theta_m) &= V_m \end{aligned}$$

---

① 注意，这里 $\hat{b}$ 依赖于样本量 n，应记为 $\hat{b}_n$. 对任何正整数 K 和 N，令

$$A_K^N = \left\{存在\ m \geqslant N\ 使得\ |\hat{b}_m - b| \geqslant \frac{1}{K}\right\}$$

$$A_K = \bigcap_{N=1}^{\infty} A_K^N, A = \bigcup_{K=1}^{\infty} A_K$$

易知 $\{\overline{\lim_{n\to\infty}} |\hat{b}_n - b| \neq 0\} \subset A$. 另一方面 $A_K \subset A_K^N \subset \bigcup_{m=N}^{\infty}\left\{|\hat{b}_m - b| \geqslant \frac{1}{K}\right\}$，$P(A_K) \leqslant \sum_{m=N}^{\infty} P\left(|\hat{b}_m - b| \geqslant \frac{1}{K}\right)$，令 $N\to\infty$ 知 $P(A_K)=0(K\geqslant 1)$. 于是 $P(A)=0$. 从而 $P(\overline{\lim_n}|\hat{b}_n - b| \neq 0)=0$. 这就证明了

$$P(\lim_n \hat{b}_n = b) = 1$$

可解出：

$$\theta_1 = f_1(V_1, \cdots, V_m)$$
$$\theta_2 = f_2(V_1, \cdots, V_m)$$
$$\cdots\cdots$$
$$\theta_m = f_m(V_1, \cdots, V_m)$$

设 $x_1, x_2, \cdots, x_n$ 是 X 的样本，用样本矩

$$\tilde{V}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{1.9}$$

来估计 $V_k(k=1,2,\cdots,m)$，然后用

$$\tilde{\theta}_k = f_k(\tilde{V}_1, \cdots, \tilde{V}_m)$$

来估计 $\theta_k(k=1,2,\cdots,m)$. 这种估计未知参数的方法叫做矩法（这里考虑的是原点矩，某些 V_k 也可用相应的中心矩 $\mu_k = E(X-EX)^k$ 代替，然后进行类似的处理，仍称矩法）.

矩法的根据是什么呢？很简单，是强大数律. 实际上，根据强大数律知 $P(\lim\limits_{n} \tilde{V}_k = V_k)=1$. 当所有 f_i 是连续函数时，必有$P(\lim\limits_{n}\tilde{\theta}_k=\theta_k)=1$.

例 1.2 设 $X\sim N(\mu,\sigma^2)$，$x_1,\cdots,x_n$ 是样本值，试求 μ,σ^2 的矩估计.

易知 $V_1=EX=\mu$，$V_2=EX^2=\sigma^2+\mu^2$，于是

$$\begin{cases}\mu=V_1\\ \sigma^2=V_2-V_1^2\end{cases} \tag{1.10}$$

用 $\tilde{V}_1=\frac{1}{n}\sum\limits_{1}^{n}x_i$ 和 $\tilde{V}_2=\frac{1}{n}\sum\limits_{1}^{n}x_i^2$ 分别估计 V_1 和 V_2，代入(1.10)得

$$\tilde{\mu}=\frac{1}{n}\sum_{1}^{n}x_i=\bar{x}$$

$$\tilde{\sigma}^2=\frac{1}{n}\sum_{1}^{n}x_i^2-\left(\frac{1}{n}\sum_{1}^{n}x_i\right)^2=\frac{1}{n}\sum_{1}^{n}(x_i-\bar{x})^2$$

对于正态分布的参数 μ,σ^2 来说，矩估计和最大似然估计是一样的，但对于其他很多分布，它们并不一样.

例 1.3 设 X 的概率密度函数是

$$f(x,\theta)=\begin{cases}\dfrac{1}{\theta}, & 0\leqslant x\leqslant\theta\\ 0, & \text{其他}\end{cases}$$

这里 θ 是未知的正数. 设 $x_1,\cdots,x_n$ 是 X 的样本,不难看出,θ 的最大似然估计 $\hat{\theta}=\max\limits_{1\leqslant i\leqslant n} x_i$. 另一方面,$EX=\dfrac{\theta}{2}$. 故 θ 的矩估计 $\tilde{\theta}=\dfrac{2}{n}\sum\limits_{1}^{n}x_i$,$\tilde{\theta}$ 与 $\hat{\theta}$ 并不一样.

例 1.4 台风可以引起内陆降雨. 下列 36 个数是 24 小时降雨量的实际观测数据(单位:mm):

31.00	2.82	3.98	4.02	9.50	4.50	11.40
10.71	6.31	4.95	5.64	5.51	13.40	9.72
6.47	10.16	4.21	11.60	4.75	6.85	6.25
3.42	11.80	0.80	3.69	3.10	22.22	7.43
5.00	4.58	4.46	8.00	3.73	3.50	6.20
0.67						

凭经验知道这种降雨量数据一般服从 Γ 分布,分布密度为

$$f(x;\alpha,\beta)=\begin{cases}0, & x\leqslant 0\\ \dfrac{1}{\alpha^{\beta}\Gamma(\beta)}x^{\beta-1}\mathrm{e}^{-x/\alpha}, & x>0\end{cases}$$

我们可用矩法估计参数 α,β. 用 X 表示降雨量;用 $x_1,x_2,\cdots,x_{36}$ 表示上述 36 个数. 易知

$$V_1=EX=\alpha\beta$$

$$V_2=EX^2=\alpha^2\beta(\beta+1)$$

$$\widetilde{V}_1=\overline{X}=\frac{1}{36}\sum_{1}^{36}x_i=7.29$$

$$\widetilde{V}_2=\frac{1}{36}\sum_{1}^{36}x_i^2=85.59$$

解方程组

$$\begin{cases}\alpha\beta=7.29\\ \alpha^2\beta(\beta+1)=85.59\end{cases}$$

得 $\tilde{\alpha}=4.45,\tilde{\beta}=1.64$.

以上介绍了最大似然估计法和矩估计法，还有许多其他的估计法. 由于估计问题多种多样，需要人们针对各个具体问题的实际情况制定恰当的估计方法，以满足实际工作的需要，读者从下面的实际例子不难理解这一点.

例 1.5 估计方法曾成功地用于情报工作. 在第二次世界大战期间，盟军曾用一种"序列号"的统计方法估计德国的武器生产，结果非常有效. 事情是这样的：德国的每件装备(无论是 V—2 火箭、坦克还是汽车轮胎等)上都刻有一个数字(序列号)，表明它按该类装备制造出来的次序排在第几. 例如，到某天为止制造出的马克—I 坦克共有 N 辆，每辆坦克带有一个号码(从 1 至 N 这 N 个整数中的一个). N 的值是盟军所不知道的. 在战争进行过程中，某些号码(序列号)被盟军知道了(由于某些德军坦克被俘获或者德军的邮件被截获)，例如 $k_1<k_2<\cdots<k_n$ 是盟军获知的 $n(n\geqslant 2)$ 个序列号，问：如何估计坦克的总数 N?

我们可把这个 N 看成某个随机向量的概率分布中的参数. 从 $1,2,\cdots,N-1,N$ 这 N 个数中随机地抽取 n 个，共有 $\binom{N}{n}$ 种结果. 我们假定各种结果出现的机会均等. 出现 $(k_1,k_2,\cdots,k_n)$ 的概率为 $\binom{N}{n}^{-1}$. N 究竟比 k_n(n 个序列号中的最大者)大多少呢? 自然想到应该利用 $k_1<k_2<\cdots<k_n$ 中相邻两个数的间隔来估计. 一种办法是用

$$W_1\overset{\mathrm{d}}{=}k_n+\frac{1}{n-1}\sum_{i=2}^{n}(k_i-k_{i-1}-1) \qquad \text{(参看[10])}$$

来估计 N. 不难看出

$$W_1=k_n+\frac{1}{n-1}(k_n-k_1)-1 \tag{1.11}$$

可以证明,用 W_1 估计 N 是没有系统偏差的. 也就是说,虽然 k_1, $k_2, \cdots, k_n$ 的值有随机性,但是 W_1 的数学期望恰好是 N.

注 我们可以证明

$$EW_1 = N \tag{1.12}$$

实际上,对任何 $i(1 \leqslant i \leqslant N-n+1)$ 有

$$P(k_1 = i) = \binom{N-i}{n-1} \cdot \binom{N}{n}^{-1} \tag{1.13}$$

对任何 $i(n \leqslant i \leqslant N)$ 有

$$P(k_n = i) = \binom{i-1}{n-1} \cdot \binom{N}{n}^{-1} \tag{1.14}$$

$$\begin{aligned} \text{于是 } Ek_n &= \sum_{i=n}^{N} iP(k_n = i) \\ &= n\sum_{i=n}^{N} \binom{i}{n} \cdot \binom{N}{n}^{-1} \\ &= n\binom{N+1}{n+1}\binom{N}{n}^{-1} = \frac{n}{n+1}(N+1) \end{aligned} \tag{1.15}$$

这里用到了恒等式:

$$\sum_{i=n}^{N} \binom{i}{n} = \binom{N+1}{n+1}$$

(这个恒等式利用关系式 $\binom{i}{n} + \binom{i}{n+1} = \binom{i+1}{n+1}$ 容易证明).

$$\begin{aligned} Ek_1 &= \sum_{i=1}^{N-n+1} iP(k_1 = i) \\ &= \sum_{i=1}^{N-n+1} i\binom{N-i}{n-1}\binom{N}{n}^{-1} \\ &= \sum_{l=n-1}^{N-1} (N-l)\binom{l}{n-1} \cdot \binom{N}{n}^{-1} \\ &= \left[(N+1)\sum_{l=n-1}^{N-1}\binom{l}{n-1} - \sum_{l=n-1}^{N-1}(l+1)\binom{l}{n-1}\right]\binom{N}{n}^{-1} \\ &= \left[(N+1)\sum_{l=n-1}^{N-1}\binom{l}{n-1} - n\sum_{k=n}^{N}\binom{k}{n}\right]\binom{N}{n}^{-1} \\ &= \left[(N+1)\binom{N}{n} - n\binom{N+1}{n+1}\right]\binom{N}{n}^{-1} = \frac{N+1}{n+1} \end{aligned}$$

由此可见,

$$
\begin{aligned}
EW_1 &= \frac{n}{n-1}Ek_n - \frac{1}{n-1}Ek_1 - 1 \\
&= \frac{n^2(N+1)}{(n-1)(n+1)} - \frac{N+1}{(n-1)(n+1)} - 1 \\
&= N
\end{aligned}
$$

根据这个估计方法,如果有 5 辆坦克被俘获,所带的序列号是 14,28,92,146,298,则 N 的估计值是

$$W_1 = 298 + \frac{284}{4} - 1 = 368.$$

第二次世界大战结束后,盟军查阅了德国的官方记录,发现上述序列号方法得到的估计远比使用其他信息来源作出的估计精确得多. 例如采用上述序列号方法估计出德国 1942 年生产坦克3 400 辆,很接近实际情况,而盟军利用间谍活动及其他情报推测出的数字是 18 000 辆,与实际情况大相径庭. 出现这种情况的原因是许多情报受纳粹德国的夸大宣传的影响,序列号方法则比较客观.

郑忠国(1938—)告诉作者,估计量 W_1(见(1.11))还可以改进,他指出估计量

$$W_2 = k_n + \frac{1}{n}\sum_{i=1}^{n}(k_i - k_{i-1} - 1) \quad (k_0 \overset{\mathrm{d}}{=\!=} 0)$$

的数学期望也是 N,而且 W_2 的方差最小.

不难看出

$$W_2 = \frac{n+1}{n}k_n - 1 \tag{1.16}$$

注 从(1.15)和(1.16)知

$$EW_2 = N \tag{1.17}$$

注意 $k_1, k_2, \cdots, k_n$ 都是随机变量,它们的概率分布与 N 有关. N 不同概率分布也就变了. 因而(1.12)和(1.17)中的数学期望符号与 N 有关,严格说应用 E_N 代替 E. (1.12)和(1.17)的准确含义是:对一切 $N \geqslant n$,等号成立.

我们可用初等方法直接证明下列重要结论:设 $\psi(k_1, \cdots, k_n)$ 是 N 的任何估计,只要满足 $E_N\psi(k_1, \cdots, k_n) = N$(对一切 $N \geqslant n$),就一定成立

$$E_N(\psi(k_1,\cdots,k_n)-N)^2 \geqslant E_N(W_2-N)^2 \tag{1.18}$$

而且当 $\psi(k_1,\cdots,k_n) \neq W_2$ 时(1.18)中左端大于右端.

证明稍长. 不妨设(1.18)中左端有限. 令

$$\begin{aligned}\tilde{\psi}(l) &\stackrel{\text{d}}{=\!=} E_N(\psi(k_1,\cdots,k_n) \mid k_n=l) \\ &= \sum_{i_1<\cdots<i_n=l} \psi(i_1,\cdots,i_n)P(k_1=i_1,\cdots,k_n=i_n \mid k_n=l) \\ &= \sum_{i_1<\cdots<i_n=l} \psi(i_1,\cdots,i_n)\binom{l-1}{n-1}\end{aligned}$$

则 $$\begin{aligned}E_N(\psi-N)^2 &= E_N(\psi-\tilde{\psi}(k_n)+\tilde{\psi}(k_n)-N)^2 \\ &= E_N(\psi-\tilde{\psi}(k_n))^2+E_N(\tilde{\psi}(k_n)-N)^2 \\ &\quad +2E_N(\psi-\tilde{\psi}(k_n))(\tilde{\psi}(k_n)-N)\end{aligned} \tag{1.19}$$

我们指出

$$E_N\psi \cdot \tilde{\psi}(k_n) = E_N(\tilde{\psi}(k_n))^2$$

实际上,$E_N\psi \cdot \tilde{\psi}(k_n) = \sum_{l=n}^{N} P(k_n=l)E_N(\psi\tilde{\psi}(k_n) \mid k_n=l) = \sum_{l=n}^{N} P(k_n = l)\tilde{\psi}(l)E_N(\psi \mid k_n=l) = \sum_{l=n}^{N} \tilde{\psi}(l)P(k_n=l)\tilde{\psi}(l) = E_N(\tilde{\psi}(k_n))^2$

另一方面,$$\begin{aligned}E_N\tilde{\psi}(k_n) &= \sum_{l=n}^{N} P(k_n=l)\tilde{\psi}(l) \\ &= \sum_{l=n}^{N} P(k_n=l)E_N(\psi(k_1,\cdots,k_n) \mid k_n=l) \\ &= E_N\psi(k_1,\cdots,k_n) = N\end{aligned}$$

于是(1.19)中右端第三项是 0. 下面指出

$$\tilde{\psi}(k_n) = W_2 \tag{1.20}$$

实际上,因为 $E_N(\tilde{\psi}(k_n)-W_2)=N-N=0$(对一切 $N \geqslant n$),即

$$\sum_{l=n}^{N}\left[\tilde{\psi}(l)-\left(\frac{n+1}{n}l-1\right)\right]P(k_n=l)=0 \tag{1.21}$$

由于 $P(k_n=l)>0$(一切 $l=n,n+1,\cdots,N$). 令 $N=n,n+1,\cdots$ 从(1.21)知 $\tilde{\psi}(l)=\frac{n+1}{n}l-1(l=n,n+1,\cdots,N)$. 故(1.20)成立.

利用(1.19)和(1.20)知

$$E_N(\psi-N)^2=E_N(\psi-W_2)^2+E_N(W_2-N)^2$$

故(1.18)成立，且 $\psi \neq W_2$ 时 $E_N(\psi-N)^2 > E_N(W_2-N)^2$.

换句话说，W_2 是所有“无偏估计”(见下一节的定义 2.1)中方差最小的.

上述序列号方法的合理性(见(1.12)和(1.17))的证明中，我们假设 n 是非随机的，因此才有(1.13)和(1.14)成立. 有人可能认为这个假设不太合理：战争中被俘获(或击毁，下同)的敌军坦克数目应是随机的. 因此另一种假设是，共有 N 辆坦克，每辆被俘获的概率为 p. 当 N 和 p 未知时，如何由数据 $1\leqslant k_1<k_2<\cdots<k_n\leqslant N$ 来估计 N?

用 τ 表示被俘获的坦克数目，τ 是一个随机变量，$\tau=n$ 则表示 N 辆中恰俘获 n 辆. 易知，当 $1\leqslant i_1<i_2<\cdots<i_n\leqslant N$ 时

$$P(k_1=i_1,\cdots,k_n=i_n\mid\tau=n)=\binom{N}{n}^{-1}$$

于是 $$P(k_1=i\mid\tau=n)=\binom{N-i}{n-1}\binom{N}{n}^{-1}\quad(i=1,2,\cdots,N-n+1)$$

$$P(k_n=i\mid\tau=n)=\binom{i-1}{n-1}\binom{N}{n}^{-1}\quad(i=n,\cdots,N)$$

从而，和(1.13)、(1.17)、(1.18)类似，有

$$E_N(W_1\mid\tau=n)=N\quad(\text{一切 } N\geqslant n)$$

$$E_N(W_2\mid\tau=n)=N\quad(\text{一切 } N\geqslant n)$$

而且若 $\psi(k_1,\cdots,k_\tau)$ 满足 $E_N(\psi(k_1,\cdots,k_\tau)\mid\tau=n)=N$(一切 $N\geqslant n$)，则必有

$$E_N((\psi(k_1,\cdots,k_\tau)-N)^2\mid\tau=n)\geqslant E_N((W_2-N)^2\mid\tau=n).$$

换句话说，即使把被俘获的坦克数目 τ 看成随机变量，则在 $\tau=n$ 的条件下，W_1 和 W_2 都是 N 的“条件无偏估计”而且 W_2 还是条件方差最小的“条件无偏估计”.

以上再一次说明了序列号方法(用 W_1 或 W_2)是有道理的.

§2 估计的优良性标准

同一个未知参数的估计量有很多种，即使是使用最大似然估计，样本量不一样，估计量也不相同. 什么样的估计量最好呢？这就涉及估计量的优良性标准. 从直观上看，自然是估计量与被估计的量越接近越好. 但“接近”二字的含义不很简单，有深入研究

的必要.

设 X 的密度函数为 $f(x,\theta)$，其中 $\theta=(\theta_1,\cdots,\theta_m)\in\Theta$，$\Theta$ 是 R^m 中的非空集合. 设 $g(\theta)$ 是 θ 的函数，$X_1,X_2,\cdots,X_n$ 是 X 的样本. 所谓 $g(\theta)$ 的估计量，是指样本的函数 $\varphi(X_1,\cdots,X_n)$. φ 的不同选择就得到不同的估计量. 直观上看，$|\varphi(X_1,\cdots,X_n)-g(\theta)|$ 越小，φ 就越好. 但是 $\varphi(X_1,\cdots,X_n)$ 的值是依赖于样本值的，它本身是随机变量，而 $g(\theta)$ 是未知的. 所以评价估计量的优劣不很简单，需要衡量优良性的标准. 当然这种标准不是唯一的，从不同角度出发可提出不同的标准.

由于样本 $X_1,\cdots,X_n$ 的联合密度函数（或概率函数）与 θ 有关，故 $\varphi(X_1,\cdots,X_n)$ 的期望与 θ 有关，有时记为 $E_\theta\varphi(X_1,\cdots,X_n)$. $\varphi(X_1,\cdots,X_n)$ 的方差记为 $\mathrm{Var}_\theta\varphi(X_1,\cdots,X_n)$.

定义 2.1 称 $\varphi(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的无偏估计，若

$$E_\theta\varphi(X_1,\cdots,X_n)=g(\theta)\text{（一切 }\theta\in\Theta\text{）}$$

定义 2.2 设 $\varphi(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的估计，称 $M_\theta(\varphi)=E_\theta[\varphi(X_1,\cdots,X_n)-g(\theta)]^2$ 为 φ 的均方误差.

显然，均方误差越小，估计就越好. 当 φ 是无偏估计时，φ 的均方误差就变为 φ 的方差了.

定义 2.3 若 $\varphi_1(X_1,\cdots,X_n)$ 和 $\varphi_2(X_1,\cdots,X_n)$ 都是 $g(\theta)$ 的估计量，且 $M_\theta(\varphi_1)\leqslant M_\theta(\varphi_2)$（一切 $\theta\in\Theta$），则称 φ_1 不次于 φ_2，此时若还有 θ_0 使得 $M_{\theta_0}(\varphi_1)<M_{\theta_0}(\varphi_2)$，则称 φ_1 比 φ_2 有效.

例如 $g(\theta)=E_\theta X$ 时，$\overline{X}_n=\frac{1}{n}\sum_1^n X_i$ 是 $g(\theta)$ 的无偏估计且 $M_\theta(\overline{X}_n)=\mathrm{Var}_\theta(\overline{X}_n)=\frac{1}{n}\mathrm{Var}_\theta(X)$. 不难看出，只要 $0<\mathrm{Var}_\theta(X)<\infty$，则 $\overline{X}_n$ 比 $\overline{X}_{n-1}$ 有效. 又如，当 $\sum_1^n\lambda_i=1$ 时 $\varphi_1=\frac{1}{n}\sum_1^n X_i$ 与 $\varphi_2=\sum_1^n\lambda_iX_i$ 都是由样本得到的 $E_\theta X$ 的无偏估计，且 φ_1 不次于 φ_2. 若 $\lambda_1,\cdots,\lambda_n$ 不全相等，则 φ_1 比 φ_2 有效.

因为均方误差是θ的函数，两个估计量并不一定总能比较优劣. 在绝大多数情况下，不可能有不次于所有估计量的估计量，即不存在均方误差最小的估计量. 因此在确定估计的优良性标准时，往往不得不缩小范围，例如在所有无偏估计中找方差最小的.

设$X_1,\cdots,X_n$是X的样本，下述定理给出了方差$\mathrm{Var}_\theta(X)$的一个无偏估计.

定理 2.1 设X的方差$\mathrm{Var}_\theta(X)$存在，$S^2 \overset{\mathrm{d}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \left(这里\ \overline{X} = \frac{1}{n}\sum_{1}^{n} X_i\right)$，则$S^2$是$\mathrm{Var}_\theta(X)$的无偏估计，即

$$E_\theta S^2 = \mathrm{Var}_\theta(X)$$

证明：

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{1}^{n} X_i^2 - n(\overline{X})^2$$

于是

$$E_\theta S^2 = \frac{1}{n-1}\sum_{i=1}^{n} E_\theta X_i^2 - \frac{n}{n-1}E_\theta(\overline{X})^2$$

$$= \frac{n}{n-1}[E_\theta X^2 - E_\theta(\overline{X})^2]$$

$$= \frac{n}{n-1}[\mathrm{Var}_\theta(X) + (E_\theta X)^2 - \mathrm{Var}_\theta(\overline{X}) - (E_\theta \overline{X})^2]$$

$$= \frac{n}{n-1}\left[\mathrm{Var}_\theta(X) - \frac{1}{n}\mathrm{Var}_\theta(X)\right] = \mathrm{Var}_\theta(X)$$

证毕.

从强大数律知，S^2是$\mathrm{Var}_\theta(X)$的强相合估计.

定义 2.4 设$\varphi(X_1,\cdots,X_n)$是$g(\theta)$的无偏估计，且对一切无偏估计$\psi(X_1,\cdots,X_n)$均有$M_\theta(\varphi) \leqslant M_\theta(\psi)$（一切$\theta$），则称$\varphi$是$g(\theta)$的（一致）最小方差无偏估计.

“最小方差无偏估计”就是一种最优的估计量，它在很多情况下确实存在，当然也有不存在的时候. 有时无偏性标准显得不合

理(参看后面的例子). 还有别的优良性标准,如贝叶斯标准,minimax标准等等,这里不介绍了(参看第七章).

怎样得到(一致)最小方差无偏估计呢? 这不是容易的事,需要具体问题具体分析,没有一个普遍适用的公式. 但在20世纪50年代形成的 Blackwell - Lehmann - Scheffe 理论对于寻求最小方差无偏估计有重要的指导作用,这个理论的严密论述涉及测度论等较深的数学工具,这里只从方法的角度作一粗浅介绍. 为此先介绍有广泛重要意义的概念——充分统计量.

定义 2.5 设 $X_1,\cdots,X_n$ 是来自总体 $F(x;\theta)$ 的简单随机样本 $(\theta\in\Theta)$,称 $U=\varphi(X_1,\cdots,X_n)$($\varphi$ 可以是向量值的函数)是 θ 的充分统计量,若 $X_1,\cdots,X_n$ 的联合密度函数(或概率函数)$L(x_1,\cdots,x_n;\theta)$ 有分解式:

$$L(x_1,\cdots,x_n;\theta)=q[\varphi(x_1,\cdots,x_n),\theta]h(x_1,\cdots,x_n) \quad (2.1)$$

(一切 $\theta\in\Theta$),其中 $h(\cdot)$ 是不依赖于 θ 的非负函数①.

充分统计量的意义在于:它概括了样本 $x_1,\cdots,x_n$ 中关于 θ 的全部信息. 所以在估计 $g(\theta)$ 时,仅利用充分统计量即可,而不必再考虑 $(x_1,\cdots,x_n)$. 例如,若 $x_1,\cdots,x_n$ 是两点分布 $B(1,p)$ 的简单随机样本,则易知 $U=\sum_1^n x_i$ 是 θ 的充分统计量. 估计 p 时,只需知道 n 次试验中的成功次数即已足够,而不必再关心具体的哪次成功 $(x_i=1)$ 哪次失败 $(x_i=0)$. 从(2.1)看出,θ 的最大似然估计如果存在,则最大似然估计一定是充分统计量的函数.

---

① 可以证明,充分统计量有下列明确的概率意义:对任何随机事件 $\{(X_1,\cdots,X_n)\in A\}$ 及 u_0,条件概率 $P((X_1,\cdots,X_n)\in A\mid\varphi(X_1,\cdots,X_n)=u_0)$ 与 θ 无关.

实际上,有更强的结论:为了 $U=\varphi(X_1,\cdots,X_n)$ 是充分统计量,必须且只须上述条件概率与 θ 无关. 这就是有名的 Fisher-Neyman 因子分解定理. 在一般书上把上述条件概率与 θ 无关作为 φ 是充分统计量的定义. 要注意的是,上述条件概率不很简单,因为在很多情形下"$\varphi(X_1,\cdots,X_n)=u_0$"是一个零概率事件,不用测度论工具无法处理. 充分统计量的直观意义是:它概括了样本 $(X_1,\cdots,X_n)$ 提供的信息.

从上述充分统计量的定义容易寻求充分统计量.

例 2.1 指数分布.

密度函数是

$$f(x,\lambda) = I_{(0,\infty)}(x)\lambda e^{-\lambda x}, \lambda \in (0,\infty)$$

其中 $I_E(x)$是集合 E 的示性函数.

此时样本$(X_1,\cdots,X_n)$的联合密度函数为

$$\prod_1^n f(x_i,\lambda) = \prod_1^n I_{(0,\infty)}(x_i) \cdot \lambda^n e^{-\lambda \sum_1^n x_i}$$

与(2.1)式对比知，$\varphi(X_1,\cdots,X_n) = \sum_1^n X_i$ 是λ 的充分统计量.

例 2.2 正态分布.

密度函数是

$$f(x;\mu,\delta) = \frac{1}{\sqrt{2\pi\delta}} \exp\left\{-\frac{1}{2\delta}(x-\mu)^2\right\}$$

其中 $\mu \in (-\infty,\infty), \delta \in (0,\infty)$.

此时样本 $X_1,\cdots,X_n$ 的联合密度函数是

$$\begin{aligned}\prod_1^n f(x_i;\mu,\delta) &= \left(\frac{1}{\sqrt{2\pi\delta}}\right)^n \exp\left\{-\frac{1}{2\delta}\sum_1^n (x_i-\mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\delta}}\right)^n \exp\left\{-\frac{1}{2\delta}\left[\sum_1^n (x_i-\overline{x})^2 \right.\right. \\ &\qquad \left.\left. + n(\overline{x}-\mu)^2\right]\right\}\end{aligned}$$

其中 $\overline{x} = \frac{1}{n}\sum_1^n x_i$. 可见，$\varphi(X_1,\cdots,X_n) = \left(\overline{X}, \sum_1^n (X_i-\overline{X})^2\right)$ 是(μ,δ) 的充分统计量. 同理知，$\left(\sum_1^n X_i, \sum_1^n X_i^2\right)$也是$(\mu,\delta)$ 的充分统计量.

例 2.3 指数型分布.

密度函数(或概率函数)是

$$f(x,\theta) = S(\theta)h(x)\exp\left\{\sum_{k=1}^{m} C_k(\theta)T_k(x)\right\}$$

其中 $\theta=(\theta_1,\cdots,\theta_m)\in\Theta, S(\theta)>0, h(x)\geqslant 0, \Theta$ 是 m 维欧氏空间中的开集.

许多常见的分布如指数分布、二项分布、Poisson 分布、正态分布等都是指数型分布(请读者自己验证).

此时样本$(X_1,\cdots,X_n)$的联合密度函数(或概率函数)为

$$\prod_{1}^{n} f(x_i,\theta)=[S(\theta)]^n \prod_{i=1}^{n} h(x_i)\exp\left\{\sum_{k=1}^{m} C_k(\theta)\sum_{i=1}^{n} T_k(x_i)\right\}$$

可见,$\varphi(X_1,\cdots,X_n)=\left(\sum_{1}^{n} T_1(X_i),\cdots,\sum_{1}^{n} T_m(X_i)\right)$是$\theta$的充分统计量.

例 2.4 均匀分布(不属于指数型分布!).

密度函数为

$$f(x;a,b)=I_{[a,b]}(x)\frac{1}{b-a}$$

这时,样本$(X_1,\cdots,X_n)$的联合密度函数为

$$\prod_{1}^{n} f(x_i,a,b)=\left(\frac{1}{b-a}\right)^n \prod_{1}^{n} I_{[a,b]}(x_i)$$

$$=\left(\frac{1}{b-a}\right)^n I_{[a,\infty)}(\min x_i) I_{(-\infty,b]}(\max x_i)$$

可见 $\varphi(X_1,\cdots,X_n)=(\min X_i,\max X_i)$是$(a,b)$的充分统计量.

充分统计量概括了样本中所含未知参数的信息. 充分统计量不止一个,例如样本本身$(X_1,\cdots,X_n)$就是未知参数的一个充分统计量(这从(2.1)直接看出). 什么样的充分统计量最有价值呢? 显然,充分统计量 $\varphi=(\varphi_1(X_1,\cdots,X_n),\cdots,\varphi_l(X_1,\cdots,X_n))$的维数 l 越小就越有价值,因为我们用尽可能少的量概括了样本中提供的信息. 例如设样本值 $x_1,\cdots,x_n$ 来自正态分布,我们可用两个量 $\overline{x}$, $\frac{1}{n}\sum_{1}^{n}(x_i-\overline{x})^2$ 来概括 n 个原始数据提供的信息. 以上的几个

例子里都存在降维（即维数小于样本量）的充分统计量；而且不管样本量多么大，永远有维数与未知参数个数相等的充分统计量，这当然是好事.

但要注意的是，并不是任何分布族永远有降维的充分统计量.例如著名的韦布尔分布，其密度函数为

$$f(x;m,\eta)=I_{(0,\infty)}(x)\eta^{-m}\cdot mx^{m-1}\exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}$$

$(m>0,\eta>0)$. 可以证明，当 $n\geqslant 3$ 时任何统计量 $\varphi=(\varphi_1(X_1,\cdots,X_n),\cdots,\varphi_l(X_1,\cdots,X_n))(l<n)$ 都不是 (m,η) 的充分统计量.

还有一个重要概念是所谓完全性. 称统计量 $\varphi(X_1,\cdots,X_n)$ 是完全的，若对任何（Borel 可测）函数 $u(\cdot)$，只要 $E_\theta u[\varphi(X_1,\cdots,X_n)]=0$（对一切 θ）就可推出 $P_\theta(u[\varphi(X_1,\cdots,X_n)]=0)=1$（对一切 θ）. 这里 P_θ 是与参数 θ 相应的概率，E_θ 是与 P_θ 相应的数学期望.

可以证明，前面几个例子中给出的充分统计量都是完全的. 例如，若 $X_1,\cdots,X_n$ 是两点分布 $B(1,p)$ 的简单随机样本，则 $\varphi=\sum_1^n X_i$ 是完全的. 实际上，因为 φ 服从二项分布 $B(n,p)$，对任何函数 $u(\cdot)$，如果 $E_p u(\varphi)=0$（一切 $p\in(0,1)$），则 $\sum_{k=0}^n u(k)C_n^k p^k(1-p)^{n-k}=0$（一切 $p\in(0,1)$）. 令 $y=p/(1-p)$，则对任何 $y\in(0,\infty)$ 有 $\sum_{k=0}^n u(k)C_n^k y^k=0$. 但此式等号左端是 y 的多项式（次数不超过 n），故必有 $u(k)=0(k=0,1,\cdots,n)$. 即 φ 是完全的.

可以证明，若参数 θ 的集合 Θ 有内点，则指数型分布族中的充分统计量 $\varphi=\left(\sum_1^n T_1(x_i),\cdots,\sum_1^n T_m(x_i)\right)$ 是完全的.

下面不加证明地叙述 Blackwell－Lehmann－Sheffe 定理. 它的证明涉及较深的数学工具，需要条件期望等概念.

Blackwell－Lehmann－Sheffe 定理：若 $\varphi(X_1,\cdots,X_n)$ 是完全的

充分统计量，$\psi[\varphi(X_1,\cdots,X_n)]$是$g(\theta)$的无偏估计，则$\psi[\varphi(X_1,\cdots,X_n)]$就是$g(\theta)$的最小方差无偏估计.

利用这个定理可以得到许多有意义的结果. 例如对正态分布$N(\mu,\sigma^2)$而言，样本均值$\overline{X}$是μ的最小方差无偏估计，样本方差$S^2=\frac{1}{n-1}\sum_1^n(X_i-\overline{X})^2$是$\sigma^2$的最小方差无偏估计. 在可靠性工程实践中有下列问题：设$X\sim N(\mu,\sigma^2)$，μ,σ^2未知，如何根据样本值$x_1,\cdots,x_n$找出概率$P(X>c)$的最小方差无偏估计？经过Kolmogorov的深刻研究(1950)，得到$P(X>c)$的最小方差无偏估计$\psi(x_1,\cdots,x_n)$如下：

$$\psi(x_1,\cdots,x_n)=\int_a^\infty p(u)\mathrm{d}u(n>2)$$

其中

$$a=\frac{\sqrt{n}(c-\overline{x})}{\sqrt{n-1}\cdot\sqrt{\sum_1^n(x_i-\overline{x})^2}}\quad\left(\overline{x}=\frac{1}{n}\sum_1^n x_i\right)$$

$$p(u)=\begin{cases}0, & |u|>1\\ \frac{1}{B\left(\frac{1}{2},\frac{n-2}{2}\right)}(1-u^2)^{\frac{n-4}{2}}, & |u|\leqslant 1\end{cases}$$

$$B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x$$

对于两总体情形：$X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$且$X$与$Y$独立，如何估计$P(X>Y)$？这是工程结构可靠性研究中很重要的问题. 设$X$的样本是$X_1,\cdots,X_n$，$Y$的样本是$Y_1,\cdots,Y_m$，当$n=m$时$X_1-Y_1,\cdots,X_n-Y_n$是$X-Y$的样本，可用上面介绍的公式来估计$P(X-Y>0)$(即$P(X>Y)$)；当$n\neq m$时，如何找出$P(X>Y)$的最小方差无偏估计？这方面也有估计公式，是20世纪70年代初得到的. 公式较繁，从略.

关于无偏估计，我们指出它有时不存在.

例 2.5 设 X 服从二项分布 $B(n,\theta)(0\leqslant\theta\leqslant1)$，即

$$P(X=k)=\binom{n}{k}\theta^k(1-\theta)^{n-k}\quad(k=0,1,\cdots,n)$$

我们指出 $g(\theta)=\dfrac{1}{1+\theta^2}$ 没有无偏估计.

实际上，若 $g(X)$ 是 $g(\theta)$ 的无偏估计（现在样本量等于 1)，则 $E_\theta\hat{g}(X)\equiv g(\theta)$（一切 θ）. 于是 $\sum_0^n g(k)C_n^k\theta^k(1-\theta)^{n-k}\equiv\dfrac{1}{1+\theta^2}(0\leqslant\theta\leqslant1)$. 此式左端是 θ 的多项式，而右端却不是，不可能在 $[0,1]$ 上恒等. 这个矛盾表明 $g(\theta)$ 没有无偏估计.

我们强调指出，不要把无偏性要求看得过分重要. 无偏性是大量重复使用同一估计量时应该尽量满足的要求，根据现有数据进行一次性估计不必要求什么无偏性. 有时无偏性的要求会导致很不合理的结果.

例 2.6 设 X 服从 Poisson 分布，即

$$P(X=k)=\frac{1}{k!}\theta^k e^{-\theta}(k=0,1,\cdots,\theta>0)$$

目标是估计 $g(\theta)=e^{-2\theta}$. 设 $g(X)$ 是 $g(\theta)$ 的无偏估计（样本量等于 1)，则

$$\sum_{k=0}^{\infty}g(k)\,\frac{1}{k!}\theta^k e^{-\theta}=e^{-2\theta}\quad(\theta\in(0,\infty))$$

于是

$$\sum_{k=0}^{\infty}\hat{g}(k)\,\frac{1}{k!}\theta^k=\sum_{k=0}^{\infty}(-1)^k\,\frac{1}{k!}\theta^k\quad(\theta>0)$$

从而 $g(k)=(-1)^k(k=0,1,\cdots)$. 这就是说，$g(\theta)$ 的无偏估计只有一个：当观测值是偶数时用 1 估计 $e^{-2\theta}$，当观测值是奇数时用 -1 估计 $e^{-2\theta}$. 这个估计量显然是不合理的，因为 $g(\theta)>0$，可是估计量 $\hat{g}(X)$ 有很大概率取值 -1.

对于无偏估计来说，当然是方差越小越好. 可是当样本量一定时，估计量的方差不能任意小，而是有下列著名的下界不等式.

定理 2.2(Cramér-Rao 不等式)　设 X 的密度函数是 $f(x,\theta)$，这里 θ 是参数，$\theta\in(a,b)(-\infty\leqslant a<b\leqslant\infty)$. $X_1,\cdots,X_n$ 是 X 的样本，$\psi(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的一个无偏估计，且满足下列正则性条件：

(1) $E\xlongequal{\mathrm{d}}\{x: f(x,\theta)\neq 0\}$ 与 θ 无关.

(2) $g'(\theta)$ 和 $\dfrac{\mathrm{d}f(x,\theta)}{\mathrm{d}\theta}$ 都存在且对一切 θ 有

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}f(x,\theta)}{\mathrm{d}\theta}\mathrm{d}x=0 \tag{2.2}$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\prod_{1}^{n}f(x_i,\theta)\right\}\mathrm{d}\underset{\sim}{x}=0 \tag{2.3}$$

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\psi(\underset{\sim}{x})\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\underset{\sim}{x}\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\psi(\underset{\sim}{x})\frac{\mathrm{d}}{\mathrm{d}\theta}\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\underset{\sim}{x}\end{aligned} \tag{2.4}$$

(这里及下面恒记 $\underset{\sim}{x}=(x_1,\cdots,x_n)$，$\mathrm{d}\underset{\sim}{x}=\mathrm{d}x_1\cdots\mathrm{d}x_n$).

(3) $I(\theta)\xlongequal{\mathrm{d}}\displaystyle\int_E\left(\frac{\mathrm{d}\ln f(x,\theta)}{\mathrm{d}\theta}\right)^2 f(x,\theta)\mathrm{d}x>0$

则有下列结论：

$$\mathrm{Var}_\theta(\psi(X_1,\cdots,X_n))\geqslant\frac{[g'(\theta)]^2}{nI(\theta)} \tag{2.5}$$

证明：注意条件(2)表示积分号与微商号可交换次序(注意 $\displaystyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\underset{\sim}{x}\equiv 1, n\geqslant 1$).

当 $I(\theta)=\infty$ 或者 $\mathrm{Var}_\theta(\psi(X_1,\cdots,X_n))=\infty$ 时(2.5)显然成立. 以下不妨设 $I(\theta)<\infty$ 且 $\mathrm{Var}_\theta(\psi)<\infty$. 由无偏性知 $g(\theta)=\displaystyle\int_E\cdots\int_E\psi(\underset{\sim}{x})\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\underset{\sim}{x}$，从条件(2.4)知

$$g'(\theta)=\int_E\cdots\int_E\psi(\underset{\sim}{x})\frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\prod_{1}^{n}f(x_i,\theta)\right\}\mathrm{d}\underset{\sim}{x}$$

$$=\int_E\cdots\int_E\psi(\underset{\sim}{x})\frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\exp\left[\sum_1^n\ln f(x_i,\theta)\right]\right\}\mathrm{d}\underset{\sim}{x}$$

$$=\int_E\cdots\int_E\psi(\underset{\sim}{x})\prod_1^n f(x_i,\theta)\cdot\sum_1^n\frac{\mathrm{d}}{\mathrm{d}\theta}\{\ln f(x_i,\theta)\}\mathrm{d}\underset{\sim}{x}$$

另一方面,从(2.3)知

$$\int_E\cdots\int_E\prod_1^n f(x_i,\theta)\cdot\sum_1^n\frac{\mathrm{d}}{\mathrm{d}\theta}\{\ln f(x_i,\theta)\}\mathrm{d}\underset{\sim}{x}=0$$

于是

$$g'(\theta)=\int_E\cdots\int_E[\psi(\underset{\sim}{x})-g(\theta)]\left[\sum_1^n\frac{\mathrm{d}}{\mathrm{d}\theta}\{\ln f(x_i,\theta)\}\right]\prod_1^n f(x_i,\theta)\mathrm{d}\underset{\sim}{x}$$

利用 Schwarz 不等式知

$$[g'(\theta)]^2\leqslant\int_E\cdots\int_E[\psi-g(\theta)]^2\prod_1^n f(x_i,\theta)\mathrm{d}\underset{\sim}{x}$$
$$\times\int_E\cdots\int_E\left\{\sum_1^n\frac{\mathrm{d}}{\mathrm{d}\theta}[\ln f(x_i,\theta)]\right\}^2\prod_1^n f(x_i,\theta)\mathrm{d}\underset{\sim}{x}$$
$$=\mathrm{Var}_\theta(\psi)\cdot E_\theta\left(\sum_1^n Y_i\right)^2 \tag{2.6}$$

其中

$$Y_i=h(X_i),\quad h(x)=\frac{\mathrm{d}}{\mathrm{d}\theta}[\ln f(x,\theta)]$$

从(2.2)知 $E_\theta Y_i=0$,从条件(3)知 $E_\theta Y_i^2=I(\theta)$,又 $Y_1,\cdots,Y_n$ 相互独立,故 $E_\theta\left(\sum_1^n Y_i\right)^2=nI(\theta)$. 从(2.6)直接推知(2.5)成立.

证毕.

不等式(2.5)就是著名的 C-R 不等式. $I(\theta)$叫做 Fisher 信息量. 对于离散型随机变量也有类似的不等式,只不过在推导过程中要用求和号代替积分号.

达到不等式(2.5)右端下界的无偏估计最引人重视,在一些较

早的书刊中称之为有效估计. 显然,如果它存在,一般就是最小方差无偏估计. 但是它不一定存在. 换句话说,C-R 不等式给出的下界有时过小.

例 2.7 设 $X \sim N(\mu, \sigma^2)$, σ^2 已知, μ 未知. $X_1, \cdots, X_n$ 是样本. 记 X 的密度为 $f(x;\mu,\sigma^2)$, 易知

$$\frac{\partial \ln f(x,\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu).$$

于是
$$I(\mu) = \frac{1}{\sigma^4}\int_{-\infty}^{\infty}(x-\mu)^2 f(x;\mu,\sigma^2)\mathrm{d}x = \frac{1}{\sigma^2}$$

记 $\overline{X} = \frac{1}{n}\sum_1^n X_i$. 则 $E_\mu \overline{X} = \mu$, $\mathrm{Var}_\mu(\overline{X}) = \frac{\sigma^2}{n} = \frac{1}{nI(\mu)}$. 可见 $\overline{X}$ 的方差达到了 C-R 不等式给出的下界.

可以证明:若 μ 已知,则 σ^2 的估计量 $\frac{1}{n}\sum_1^n (X_i-\mu)^2$ 的方差达到 C-R 不等式给出的下界.

例 2.8 设 X 服从伯努利分布,即

$$P(X=1) = 1 - P(X=0) = p \quad (0 < p < 1).$$

此时 X 的概率函数为

$$f(x,p) = p^x(1-p)^{1-x} \quad (x=0,1).$$

易知 $\frac{\mathrm{d}}{\mathrm{d}p}(\ln f(x,p)) = \frac{x}{p} - \frac{1-x}{1-p}$, 于是

$$I(p) = \sum_{x=0}^{1}\left\{\frac{\mathrm{d}}{\mathrm{d}p}\ln f(x,p)\right\}^2 f(x,p) = \frac{1}{p(1-p)}$$

我们知道, $\overline{X} = \frac{1}{n}\sum_1^n X_i$ 是 p 的无偏估计,易见

$$\mathrm{Var}_p(\overline{X}) = \frac{1}{n}\mathrm{Var}_p(X) = \frac{1}{n}p(1-p) = \frac{1}{nI(p)}$$

故 $\overline{X}$ 的方差达到了 C-R 不等式给出的下界.

例 2.9 设 X 服从 Poisson 分布,即

$$P(X=k) = \frac{1}{k!}\lambda^k \mathrm{e}^{-\lambda} \quad (k=0,1,\cdots)$$

其中 λ 是未知的正数. 我们要估计 $g(\lambda)=\mathrm{e}^{-\lambda}$.

取 $g(\lambda)$ 的无偏估计

$$\hat{g}=\left(1-\frac{1}{n}\right)^{\sum_{1}^{n}X_i}$$

($X_1,\cdots,X_n$ 是样本). 它的方差为 $\mathrm{e}^{-2\lambda}(\mathrm{e}^{\frac{\lambda}{n}}-1)$. 由前面介绍过的 Blackwell - Lehmann - Scheffe 定理, $\hat{g}$ 是 $g(\lambda)$ 的最小方差无偏估计. 但 C-R 不等式给出的下界为 $\frac{\lambda}{n}\mathrm{e}^{-2\lambda}$, 它小于 $\mathrm{Var}_\lambda(\hat{g})$, 这说明 C-R 不等式给出的下界不可能达到.

在统计学中常常要考察统计量的大样本性质, 即当 n 很大时统计量 $\varphi(X_1,\cdots,X_n)$ 的性质. 这一点在 §1 中已提到过, 那里说的相合性就是大样本性质. 相合性的一般定义如下.

定义 2.6 称估计量 $\varphi_n(X_1,\cdots,X_n)(n\geqslant 1)$ 是 $g(\theta)$ 的相合估计, 若对任何 $\varepsilon>0$ 有

$$\lim_n P(|\varphi_n-g(\theta)|\geqslant\varepsilon)=0$$

定义 2.7 称估计量 $\varphi_n(X_1,\cdots,X_n)(n\geqslant 1)$ 是 $g(\theta)$ 的强相合估计, 若

$$P(\lim_n\varphi_n=g(\theta))=1$$

当然, 强相合估计一定是相合估计. 不难看出, 任何有意义的估计应具有相合性.

从强大数律知, 矩估计一般是有强相合性的.

最大似然估计在十分广泛的条件下也是有强相合性的, 参看文献[2].

现在就单参数情形给出一个定理.

定理 2.3 设 X 有密度函数 $f(x,\theta)$, 其中 $\theta\in(a,b)(-\infty\leqslant a<b\leqslant\infty)$, 且满足下列条件:

(1) 对一切 $\theta_1\neq\theta_2$,

$$\mu\{x: f(x,\theta_1)>0 \text{ 且 } f(x,\theta_1)\neq f(x,\theta_2)\}>0$$

这里 μ 表示 Lebesgue 测度.

(2) 似然函数 $L(x_1,\cdots,x_n;\theta)=\prod_1^n f(x_i,\theta)$ 是 θ 的单峰函数，即存在 $\hat{\theta}_n=\hat{\theta}_n(x_1,\cdots,x_n)$ 使得 $L(x_1,\cdots,x_n;\theta)$ 在$(a,\hat{\theta}_n]$上严格增，在$[\hat{\theta}_n,b)$上严格减.

则 θ 的最大似然估计 $\hat{\theta}_n$（基于样本 $X_1,\cdots,X_n$）存在且唯一，而且 $\hat{\theta}_n$ 是 θ 的强相合估计.

证明：从似然函数的单峰性假定知，最大似然估计 $\hat{\theta}_n=\hat{\theta}_n(X_1,\cdots,X_n)$存在且唯一. 下面证明它具有强相合性. 为了进行严密的讨论，我们假设 X_1，X_2，…是概率空间$(\Omega,\mathscr{F},P_\theta)$$(\theta\in(a,b))$上相互独立且同分布的随机变量列，这里 Ω 是非空集合（由全体"基本事件"组成），$\mathscr{F}$是 Ω 的一些子集（随机事件）组成的 σ 代数，P_θ 是 $\mathscr{F}$上有定义的概率测度且 $P_\theta(X_1\leqslant x)=\int_{-\infty}^x f(u,\theta)\mathrm{d}u$. 与 θ 相应的数学期望用 E_θ 表示 .

任意固定 $\theta_0\in(a,b)$. 取 $\delta_m>0(m=1,2,\cdots)$满足$\lim\limits_m\delta_m=0$ 且 $\theta_0\pm\delta_m\in(a,b)$. 利用不等式 $\ln x<x-1(x>0$ 且 $x\neq1)$及条件(1)知

$$-\infty\leqslant\int_{-\infty}^{\infty}f(x,\theta_0)\ln\frac{f(x,\theta_0\pm\delta_m)}{f(x,\theta_0)}\mathrm{d}x<0$$

$$-\infty\leqslant E_{\theta_0}\left[\ln\frac{f(X_1,\theta_0\pm\delta_m)}{f(X_1,\theta_0)}\right]<0$$

从强大数律知

$$P_{\theta_0}\left(\lim_n\frac{1}{n}\sum_{i=1}^n\ln\frac{f(X_i,\theta_0\pm\delta_m)}{f(X_i,\theta_0)}=E_{\theta_0}\ln\frac{f(X_1,\theta_0\pm\delta_m)}{f(X_1,\theta_0)}\right)=1$$

于是有

$$P_{\theta_0}\left\{\lim_n\frac{\ln L(X_1,\cdots,X_n;\theta_0\pm\delta_m)-\ln L(X_1,\cdots,X_n;\theta_0)}{n}<0\right\}=1$$

故有集合 Λ_m，$P_{\theta_0}(\Lambda_m)=0$，当 $\omega\overline{\in}\Lambda_m$ 时

$$\lim_n\frac{\ln L(X_1,\cdots,X_n;\theta_0\pm\delta_m)-\ln L(X_1,\cdots,X_n;\theta_0)}{n}<0$$

（注意 $X_1,\cdots,X_n$ 都是ω 的函数，$\omega\in\Omega$).

于是有 $n_m(\omega)$，当 $n\geqslant n_m(\omega)$时有

$$L(X_1,\cdots,X_n;\theta_0\pm\delta_m)<L(X_1,\cdots,X_n;\theta_0)\tag{2.7}$$

令 $\Lambda=\bigcup\limits_{m=1}^{\infty}\Lambda_m$，当然 $P_{\theta_0}(\Lambda)=0$. 对任何 $\omega\overline{\in}\Lambda$，只要 $n\geqslant n_m(\omega)$则(2.7)成

立. 利用 L 的单峰性知最大似然估计 $\hat{\theta}_n$ 在 $(\theta_0-\delta_m,\theta_0+\delta_m)$ 内，有 $|\hat{\theta}_n-\theta_0|<\delta_m$. 先令 $n\to\infty$，再令 $m\to\infty$，知 $\lim\limits_n\hat{\theta}_n=\theta_0$. 所以 $P_{\theta_0}(\lim\limits_n\hat{\theta}_n=\theta_0)=1$. 这就证明了 $\hat{\theta}_n$ 的强相合性. 证毕.

§3 置信区间(区间估计)

设 $x_1,\cdots,x_n$ 是样本值，前面讲的是用一个量 $\varphi(x_1,\cdots,x_n)$（估计量）去估计未知参数的函数 $g(\theta)$. 换句话说，用 $\varphi(x_1,\cdots,x_n)$ 作为 $g(\theta)$ 的近似值. 读者可能不满意，到底估计值与真值相差有多大？这是个很重要的问题，统计学家力图给出回答. 这个问题与估计量作为随机变量的概率分布有关.

定义 3.1 设 $\gamma\in(0,1)$，$\varphi_1(X_1,\cdots,X_n)$ 和 $\varphi_2(X_1,\cdots,X_n)$ 是两个统计量，$\varphi_1\leqslant\varphi_2$. 称 $[\varphi_1,\varphi_2]$ 是 $g(\theta)$ 的置信水平是 γ 的置信区间（或叫区间估计），若对一切 θ 均有

$$P\{\varphi_1(X_1,\cdots,X_n)\leqslant g(\theta)\leqslant\varphi_2(X_1,\cdots,X_n)\}\geqslant\gamma \tag{3.1}$$

置信水平也称置信度，常取 $\gamma=0.80,0.90,0.95,0.99$. 若(3.1)式的左端的下确界（对一切 θ）恰好是 γ，则称 γ 为 $[\varphi_1,\varphi_2]$ 的置信系数.

置信区间 $[\varphi_1,\varphi_2]$ 的好处在于：它以一定把握保证 $g(\theta)$ 在此区间中.

在(3.1)中的 φ_1,φ_2 分别称为 $g(\theta)$ 的置信下限、置信上限. 在某些问题里只关心置信下限，此时取 $\varphi_2=\infty$；在另外一些问题里只关心置信上限，此时取 $\varphi_1=-\infty$.

寻找置信区间是一件重要的工作，当然应该限于寻找优良的置信区间. 若不管优良性，取 $\varphi_1=-\infty,\varphi_2=\infty$，$[-\infty,\infty]$ 永远是 $g(\theta)$ 的置信水平为 γ 的置信区间. 读者一眼看出，这个置信区间毫无用处，它没有提供 $g(\theta)$ 的任何信息. 什么是优良的置信区间呢？这就涉及优良性的标准. 有多种不同的标准，我们在这里暂不进行深入的讨论，只是指出：优良的置信区间应该是区间长度

(即 $\varphi_2-\varphi_1$)比较小(如果只关心置信下限,则这种下限越大越好;如果只关心置信上限,则这种上限越小越好).

读者自然会问:为什么在定义 3.1 中(3.1)式里是"$\geqslant$"而不是"$=$"? 原因是,若要求"$=$"号处处成立,则在许多重要情况下所要求的置信区间不存在. 以后将有例子说明这一点. 当然,要力求找到置信系数恰为 γ 的置信区间(或置信限).

怎样寻找优良的置信区间呢? 这不是容易的事,要具体问题具体分析. 有三个一般性方法可指导我们对具体问题进行分析,有助于找出优良的置信区间. 本节介绍枢轴量方法和统计量方法,第三个方法是借助于假设检验理论的接受域方法,将在第三章中叙述.

首先介绍枢轴量方法,这个方法是初等统计学中最常用的,也是本书叙述的重点.

为了寻找 $g(\theta)$ 的置信区间,我们设法选择与样本 $X_1,\cdots,X_n$ 及 $g(\theta)$ 有关的函数 $h(X_1,\cdots,X_n;g(\theta))$,使得这个函数(实际是随机变量)的概率分布函数 $H(x)$ 与 θ 无关. 在此基础上找 $\lambda_1<\lambda_2$ 满足 $H(\lambda_2)-H(\lambda_1)\geqslant\gamma$. 于是 $P(\lambda_1\leqslant h(X_1,\cdots,X_n;g(\theta))\leqslant\lambda_2)\geqslant\gamma$. 解不等式 $\lambda_1\leqslant h(x_1,\cdots,x_n;u)\leqslant\lambda_2$,得到 $\varphi_1(x_1,\cdots,x_n)\leqslant u\leqslant\varphi_2(x_1,\cdots,x_n)$. 于是 $[\varphi_1(X_1,\cdots,X_n),\varphi_2(X_1,\cdots,X_n)]$ 便是 $g(\theta)$ 的置信水平为 γ 的置信区间. 当然,λ_1,λ_2 还应选得使这个置信区间的长度尽可能的小.

上述的 $h(X_1,\cdots,X_n;g(\theta))$ 一般称为枢轴量(Pivotal),它的精确分布与未知参数无关,如何找到合适的枢轴量就是问题的关键. 我们通过几个具体问题加以讨论,希望读者从中理解寻求枢轴量和置信区间的一般步骤. 至于所得到的置信区间的优良性,则不去讨论. 在第三章中将会间接地看出,这些置信区间是比较优良的,区间是"比较短的".

(一) 指数分布

密度函数是 $f(x,\lambda)=I_{(0,\infty)}(x)\lambda e^{-\lambda x}$,其中 λ 是未知的正数.

如何从样本 $X_1,\cdots,X_n$ 找出 λ 的置信区间?

从 §1 知 λ 的最大似然估计是 $n\Big/\sum_1^n X_i$. 可以求出 $\sum_1^n X_i$ 的密度函数为

$$q(x,\lambda)=I_{(0,\infty)}(x)\frac{1}{(n-1)!}\lambda^n x^{n-1}\mathrm{e}^{-\lambda x}$$

(读者可用数学归纳法验证).

令 $h=2\lambda\sum_1^n X_i$, 则 h 的密度函数是

$$p(x)=\begin{cases}\dfrac{1}{2^n\Gamma(n)}x^{n-1}\mathrm{e}^{-\frac{1}{2}x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

这表明 h 是一个枢轴量,它可用来找 λ 的置信区间.

定义 3.2 称随机变量 ξ 服从 n 个自由度的 χ^2 分布,若它的密度函数是这样的:

$$p_n(x)=\begin{cases}0, & x\leqslant 0\\ \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{1}{2}x}, & x>0\end{cases}\tag{3.2}$$

其中 $\Gamma(\alpha)=\int_0^\infty u^{\alpha-1}\mathrm{e}^{-u}\mathrm{d}u\,(\alpha>0)$ 是 Γ 函数.

χ^2 分布是统计学中最重要的分布之一,用途很广,它的分位数表是一个常用的统计表①. 不难看出,上面的枢轴量 h 正好服从 $2n$ 个自由度的 χ^2 分布. 查本书的附表 3 可找到 $\lambda_1<\lambda_2$ 满足 $P\left(\lambda_1\leqslant 2\lambda\sum_1^n X_i\leqslant\lambda_2\right)=\gamma$, 于是

---

① 设随机变量 X 的分布函数是 $F(x)$. 给定 $0<p<1$. 设 x_p 满足 $P(X\leqslant x_p)\geqslant p\geqslant P(X<x_p)$. 则称 x_p 为 X(或 $F(x)$)的 p 分位数. 不难看出,$F(x_p)\geqslant p\geqslant F(x_p-0)$. p 分位数一定存在,但有时不唯一.

$$P\left(\frac{\lambda_1}{2\sum_1^n X_i} \leqslant \lambda \leqslant \frac{\lambda_2}{2\sum_1^n X_i}\right)=\gamma$$

这表明

$$\left[\frac{\lambda_1}{2\sum_1^n X_i}, \frac{\lambda_2}{2\sum_1^n X_i}\right]$$

是 λ 的置信区间,置信水平是 γ. 为了使这个置信区间之长度最小,应使 $\lambda_2-\lambda_1$ 最小. 可以证明 λ_1, λ_2 应满足下列两个方程

$$\int_{\lambda_1}^{\lambda_2} p_{2n}(x)\mathrm{d}x=\gamma \qquad (\lambda_1<\lambda_2)$$

$$p_{2n}(\lambda_1)=p_{2n}(\lambda_2)$$

这里 $p_{2n}(x)$ 是 $2n$ 个自由度的χ^2分布的密度函数,即 $p_{2n}(x)=\frac{1}{2^n\Gamma(n)}x^{n-1}\mathrm{e}^{-\frac{x}{2}}(x>0)$. 第一个方程保证置信水平是 γ,第二个方程保证 $\lambda_2-\lambda_1$ 达到最小值.

从这两个方程就可确定出 λ_1 和 λ_2,但在实际工作中这样做颇嫌不便. 人们通常采用"平分法",即找 λ_1 和 λ_2 满足

$$\int_0^{\lambda_1} p_{2n}(x)\mathrm{d}x=\frac{1-\gamma}{2}, \int_{\lambda_2}^{\infty} p_{2n}(x)\mathrm{d}x=\frac{1-\gamma}{2}$$

而不去追求 λ_1 和 λ_2 的最优选择.

注 我们可以证明下列一般性命题:设 $p(x)$是$(-\infty,\infty)$上非负连续函数,满足下列条件:

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x=1$$

且存在$-\infty\leqslant x_0<\xi<\infty$使得 $x\leqslant x_0$ 时 $p(x)=0$,在(x_0,ξ)上$p(x)$是严格增函数,在(ξ,∞)上 $p(x)$是严格减函数. 则有下列结论:

(1) 对任何 $\gamma\in(0,1)$,方程组

$$\int_a^b p(x)\mathrm{d}x=\gamma \quad (a<b) \tag{3.3}$$

$$p(a)=p(b) \tag{3.4}$$

有唯一解(a^*,b^*).

(2) 对任何满足(3.3)的a,b,为了$b-a$达到最小值,必须且只需$a=a^*$, $b=b^*$.

证明:对任何$a\in(x_0,\xi)$,因为$0<p(a)<p(\xi)$,故有唯一的$b=b(a)$使得$p(b)=p(a)$. 易知,这个$b(a)$是a的连续减函数. 令$\psi(a)=\int_a^{b(a)}p(x)\mathrm{d}x$. 则$\psi(a)$是连续函数,且$\lim\limits_{a\to x_0^{+0}}\psi(a)=\int_{x_0}^{\infty}p(x)\mathrm{d}x=1$, $\lim\limits_{a\nearrow\xi}\psi(a)=0$(因为$\lim\limits_{a\nearrow\xi}b(a)=\xi$). 故一定有$a^*\in(x_0,\xi)$满足$\psi(a^*)=\gamma$. 令$b^*=b(a^*)$. 则$(a^*,b^*)$满足方程(3.3)、(3.4).

若(a,b)也满足(3.3)和(3.4),易知$x_0<a<\xi$. 若$a<a^*$,则$p(a)<p(a^*)=p(b^*)>p(b)$. 于是$b>b^*$. 从而$\int_a^b p(x)\mathrm{d}x>\int_{a^*}^{b^*}p(x)\mathrm{d}x=\gamma$,这与(3.3)相矛盾. 同理,若$a>a^*$,则可推出$\int_{a^*}^{b^*}p(x)\mathrm{d}x>\gamma$,这与$a^*,b^*$满足(3.3)相矛盾. 故$a=a^*$, $b=b^*$,即方程组(3.3)—(3.4)有唯一解.

令$c=b^*-a^*$, $\phi(a)=\int_a^{a+c}p(x)\mathrm{d}x$,则$\phi'(a)=p(a+c)-p(a)$. 于是$\phi'(a^*)=p(b^*)-p(a^*)=0$. 另一方面,当$a<a^*$时若$a+c\leqslant a^*$则$p(a+c)>p(a)$(因$\xi\in(a^*,a^*+c)$),若$a<a^*<a+c$则$a+c<a^*+c$,于是$p(a+c)>p(a^*)>p(a)$,总之,当$a<a^*$时$\phi'(a)>0$. 同理知$a>a^*$时$\phi'(a)<0$. 故$a^*$是$\phi(a)$的最大值点,即$\phi(a)\leqslant\phi(a^*)$(一切$a$). 设$a$, b满足(3.3). 我们指出必有$b-a\geqslant b^*-a^*$. 用反证法,设$b-a<b^*-a^*=c$则$\int_a^b p(x)\mathrm{d}x<\int_a^{a+c}p(x)\mathrm{d}x=\phi(a)\leqslant\phi(a^*)=\gamma$. 这与$a,b$满足(3.3)相矛盾. 故$b-a\geqslant b^*-a^*$.

最后证明:对任何满足(3.3)的a,b,若$b-a\geqslant b_1-a_1$(a_1,b_1也满足(3.3)),则$a_1=a^*$, $b_1=b^*$. 实际上,令$c_1=b_1-a_1$, $\phi(a)=\int_a^{a+c_1}p(x)\mathrm{d}x$. 则$\phi(a_1)=\int_{a_1}^{b_1}p(x)\mathrm{d}x=\gamma=\int_a^b p(x)\mathrm{d}x\geqslant\int_a^{a+c_1}p(x)\mathrm{d}x=\phi(a)$. 可见$\phi(a)$在$a_1$达到最大值,故$\phi'(a_1)=0$. 即有

$$p(a_1+c_1)-p(a_1)=0.$$

这表明$p(b_1)=p(a_1)$. 由于方程组(3.3)—(3.4)的解是唯一的. 故$a_1=a^*$,

$b_1 = b^*$.

全部证毕.

(二) 正态分布.

设 $X \sim N(\mu, \sigma^2)$. 问:如何从样本 $X_1, \cdots, X_n$ 找出未知参数 μ 的置信区间和 σ^2 的置信区间?

这里可细分为三个问题:

(1) 已知 σ,找 μ 的置信区间;

(2) 未知 μ,找 σ 的置信区间;

(3) 未知 σ,找 μ 的置信区间.

我们可分别找出相应的枢轴量. 对于(1),自然想到用样本均值 $\overline{X}$ 估计 μ,可以证明 $h \stackrel{d}{=} \frac{1}{\sigma}\sqrt{n}(\overline{X} - \mu)$ 服从 $N(0,1)$(见下面的定理 3.3),从而 h 是枢轴量. 对于(2),自然想到用 $S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$ 估计 σ^2,可以证明 $\eta = (n-1)S^2 \cdot \frac{1}{\sigma^2}$ 服从 $n-1$ 个自由度的 χ^2 分布(见下面的定理 3.3),从而 η 是枢轴量. 对于(3),可以证明

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S}$$

的分布与 μ, σ 无关(见下面的定理 3.3),从而 T 是枢轴量.

为了证明上面指出的几个结论,要证明几条在统计学中应用广泛的基本定理. 先从 χ^2 分布的直观意义谈起.

定理 3.1 设 $X_1, \cdots, X_n$ 相互独立同分布,共同分布是 $N(0,1)$,则 $\xi \stackrel{d}{=} \sum_1^n X_i^2$ 服从 n 个自由度的 χ^2 分布(密度函数见(3.2)).

证明:这是初等概率论里已经证明过的结论. 证明方法是计算 ξ 的分布函数,然后就可看出 ξ 的密度函数正好是(3.2). 这里从略(参看[1]).

系 3.1 若 ξ 服从 n 个自由度的 χ^2 分布，则 $E\xi=n$.

这从定理 3.1 直接推出.

按照习惯的记法，如果 ξ 服从 n 个自由度的 χ^2 分布，则记为 $\xi\sim\chi^2(n)$.

对于 χ^2 分布的变量，有下列重要性质.

系 3.2 设 ξ 与 η 相互独立，$\xi\sim\chi^2(n)$，$\eta\sim\chi^2(m)$，则

$$\xi+\eta\sim\chi^2(n+m)$$

证明：这利用定理 3.1 甚易推出，留给读者自己完成.

定理 3.2 设 $X_1,X_2,\cdots,X_n$ 相互独立，且

$$X_i\sim N(\mu_i,\sigma^2)\quad(i=1,2,\cdots,n)$$

$\mathbf{A}=(a_{ij})$ 是 n 阶正交矩阵，

$$Y_i=\sum_{k=1}^{n}a_{ik}X_k\quad(i=1,2,\cdots,n)$$

则 $Y_1,\cdots,Y_n$ 相互独立，且

$$Y_i\sim N\left(\sum_{k=1}^{n}a_{ik}\mu_k,\sigma^2\right)\quad(i=1,2,\cdots,n)$$

证明：先考虑特殊情形：$\mu_i\equiv0\,(i=1,2,\cdots,n)$. 此时 $X_1,\cdots,X_n$ 的联合密度是

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{1}^{n}x_i^2\right\}$$

任给定 $t_1,\cdots,t_n$，令

$$D=\left\{(x_1,\cdots,x_n):\sum_{j=1}^{n}a_{ij}x_j\leqslant t_i,i=1,2,\cdots,n\right\}$$

则

$$P(Y_1 \leqslant t_1, \cdots, Y_n \leqslant t_n) = \int_D \cdots \int \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n x_i^2\right\} \mathrm{d}\underset{\sim}{x}$$

(这里 $\mathrm{d}\underset{\sim}{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n$).

作变数替换:

$$y_i = \sum_{k=1}^n a_{ik} x_k \quad (i = 1, 2, \cdots, n)$$

则

$$x_i = \sum_{k=1}^n a_{ki} y_k \quad (i = 1, 2, \cdots, n)$$

变数替换的雅可比式为

$$J\left(\frac{x_1, \cdots, x_n}{y_1, \cdots, y_n}\right) = |A'| = 1 \text{ 或 } -1$$

又

$$\sum_1^n x_i^2 = \sum_1^n y_i^2$$

故有

$$P(Y_1 \leqslant t_1, \cdots, Y_n \leqslant t_n)$$

$$= \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n y_i^2\right\} \mathrm{d}y_1 \cdots \mathrm{d}y_n$$

$$= \prod_{i=1}^n \left[\int_{-\infty}^{t_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{y_i^2}{2\sigma^2}\right\} \mathrm{d}y_i\right]$$

在此式中当 $j \neq i$ 时,令 $t_j \to \infty$,得

$$P(Y_i \leqslant t_i) = \int_{-\infty}^{t_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{y_i^2}{2\sigma^2}\right\} \mathrm{d}y_i$$

从而

$$P(Y_1 \leqslant t_1, \cdots, Y_n \leqslant t_n) = \prod_{i=1}^n P(Y_i \leqslant t_i)$$

这表明 $Y_1, \cdots, Y_n$ 相互独立且服从同一分布 $N(0, \sigma^2)$.

现在考虑一般情形. 令 $Z_i=X_i-\mu_i$，则 $Z_1,\cdots,Z_n$ 相互独立，都服从 $N(0,\sigma^2)$. 根据刚才已证明的特殊情形，知 $\sum_{k=1}^{n}a_{1k}Z_k,\cdots,\sum_{k=1}^{n}a_{nk}Z_k$ 相互独立，且

$$\sum_{k=1}^{n}a_{ik}Z_k\sim N(0,\sigma^2)$$

由于

$$Y_i=\sum_{k=1}^{n}a_{ik}Z_k+\sum_{k=1}^{n}a_{ik}\mu_k\quad(i=1,2,\cdots,n)$$

故 $Y_1,\cdots,Y_n$ 相互独立，且

$$Y_i\sim N\left(\sum_{k=1}^{n}a_{ik}\mu_k,\sigma^2\right)\quad(i=1,2,\cdots,n)$$

证毕.

定理 3.3 设 $X_1,\cdots,X_n$ 相互独立，都服从 $N(\mu,\sigma^2)$，则

(1) $\overline{X}=\frac{1}{n}\sum_{1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$

(2) $\frac{1}{\sigma^2}\sum_{1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$

(3) $\overline{X}$ 与 $\sum_{1}^{n}(X_i-\overline{X})^2$ 相互独立.

证明：取 n 阶正交矩阵 $A=(a_{ij})$ 使得第一行元素全是 $\frac{1}{\sqrt{n}}$，这样的正交矩阵当然是有的. 令

$$Y_i=\sum_{k=1}^{n}a_{ik}X_k\quad(i=1,2,\cdots,n)$$

从定理 3.2 知，$Y_1,\cdots,Y_n$ 相互独立，且 $Y_1\sim N(\sqrt{n}\mu,\sigma^2)$，$Y_i\sim N(0,\sigma^2)(i=2,\cdots,n)$. 于是

$$\overline{X}=\frac{1}{\sqrt{n}}Y_1\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

由于

$$\sum_1^n X_i^2 = \sum_1^n Y_i^2$$

$$\sum_1^n (X_i - \overline{X})^2 = \sum_1^n X_i^2 - n(\overline{X})^2$$

$$= \sum_1^n Y_i^2 - Y_1^2 = \sum_2^n Y_i^2$$

利用定理 3.1 知

$$\frac{1}{\sigma^2}\sum_1^n (X_i - \overline{X})^2$$

服从 $n-1$ 个自由度的χ^2分布．由于 $Y_1,\cdots,Y_n$ 相互独立，故 $\overline{X}$ 与 $\sum_1^n (X_i - \overline{X})^2$ 相互独立．

证毕．

顺便说一句，定理 3.3 的(3)是正态分布极深刻的性质．经过现代的深入研究知道，只有正态分布才有这个性质．换句话说，若一个总体的样本均值 $\overline{X}$ 与样本离差 $\sum_1^n (X_i-\overline{X})^2$ 相互独立，则这个总体必服从正态分布．

有了定理 3.3，我们可以找出正态分布参数 μ,σ 的置信区间．设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$的样本，从定理 3.3(1)知

$$\frac{1}{\sigma}\sqrt{n}(\overline{X}-\mu) \sim N(0,1)$$

查正态分布数值表知

$$P\left(\left|\frac{1}{\sigma}\sqrt{n}(\overline{X}-\mu)\right| \leqslant 1.96\right) = 0.95$$

于是，σ 已知时区间

$$\left[\overline{X}-1.96\frac{\sigma}{\sqrt{n}},\overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right] \tag{3.5}$$

是 μ 的置信水平为 0.95 的置信区间. 具体说来,如果做 100 次抽样(每次抽 n 个单元),则算出的 $\overline{X}$ 值大约有 95 次使得区间(3.5)包含 μ. 当然,也可能碰到这个区间不包含 μ 的情形,但出现这种情形的可能性比较小,约为 5%.

我们还要注意,置信区间的长度与样本量 n 有关. 当然希望置信区间的长度越小越好,但为此需付出代价:n 必须大. 故在实际问题里要具体分析,适当掌握,不能走极端.

例 3.1 某车间生产滚珠,从长期实践中知道,滚珠直径 X 可以认为是服从正态分布的. 从某天的产品里随机抽取 6 个,测得直径如下(单位:mm):

14.70,15.21,14.90,14.91,15.32,15.32

试估计该天产品的直径的平均值? 如果知道该天产品直径的方差是 0.05,试找出直径平均值的置信区间(置信水平为 0.95)?

易知,这 6 个数的平均值 $\overline{X}=15.06$(mm),这就是直径平均值 μ 的估计值. 利用(3.5)知 μ 的置信区间是[14.88,15.24](置信水平为 0.95).

如果 σ^2 未知,需要找 σ^2 的置信区间,从定理 3.3 知

$$W \stackrel{\mathrm{d}}{=} \frac{1}{\sigma^2}\sum_{1}^{n}(X_i-\overline{X})^2 \sim \chi^2(n-1)$$

查书末附表 3,有 λ_1,λ_2 使得

$$P(\lambda_1 \leqslant W \leqslant \lambda_2)=0.95$$

于是

$$P\left(\frac{1}{\lambda_2}\sum_{1}^{n}(X_i-\overline{X})^2 \leqslant \sigma^2 \leqslant \frac{1}{\lambda_1}\sum_{1}^{n}(X_i-\overline{X})^2\right)=0.95$$

这表明 σ^2 的置信水平为 0.95 的置信区间是

$$\left[\frac{1}{\lambda_2}\sum_{1}^{n}(X_i-\overline{X})^2,\ \frac{1}{\lambda_1}\sum_{1}^{n}(X_i-\overline{X})^2\right] \tag{3.6}$$

可类似地找出 σ^2 的置信水平为 γ 的置信区间.

现在问:如果 σ^2 未知,如何找 μ 的置信区间?

记

$$S=\sqrt{\frac{1}{n-1}\sum_{1}^{n}(X_i-\overline{X})^2}$$

从定理 3.3 知道,随机变量 $\frac{1}{\sigma}\sqrt{n}(\overline{X}-\mu)$ 与 $\frac{1}{\sigma}S$ 相互独立,又

$$\frac{1}{\sigma}\sqrt{n}(\overline{X}-\mu)\sim N(0,1) \tag{3.7}$$

$$\frac{1}{\sigma^2}\sum_{1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1) \tag{3.8}$$

于是随机变量

$$T=\frac{\sqrt{n}(\overline{X}-\mu)}{S}$$

的分布是完全确定的,与 μ,σ 无关. 显然,如果 T 的分布函数能求出来,则可找到 λ 使得

$$P(|T|\leqslant\lambda)=\gamma\quad(0<\gamma<1) \tag{3.9}$$

于是

$$P\left(\overline{X}-\lambda\frac{S}{\sqrt{n}}\leqslant\mu\leqslant\overline{X}+\lambda\frac{S}{\sqrt{n}}\right)=\gamma$$

这表明 μ 的置信水平为 γ 的置信区间是:

$$\left[\overline{X}-\lambda\frac{S}{\sqrt{n}},\overline{X}+\lambda\frac{S}{\sqrt{n}}\right] \tag{3.10}$$

T 的分布函数是什么呢? T 服从所谓 $n-1$ 个自由度的 t 分布. t 分布的定义是这样的:

定义 3.3 称随机变量 ζ 服从 n 个自由度的 t 分布(简记为 $\zeta\sim t(n)$),若 ζ 的分布密度是

$$p_n(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{3.11}$$

t 分布的直观意义与重要性在于下列事实.

定理 3.4 设 ξ 与 η 相互独立，$\xi \sim N(0,1)$，$\eta \sim \chi^2(n)$，则

$$\zeta \stackrel{\mathrm{d}}{=} \xi \bigg/ \sqrt{\frac{1}{n}\eta}$$

服从 n 个自由度的 t 分布.

证明：这个定理在初等概率论里已证明过. 证明方法是计算 ζ 的分布函数，然后知道 ζ 的分布密度正好是(3.11)，这里从略(参看[1]). 从(3.7)，(3.8)及定理 3.4 知随机变量 T 服从 $n-1$ 个自由度的 t 分布.

t 分布的密度函数是 x 的偶函数. 当自由度 $n \geqslant 25$ 时 t 分布的密度曲线与标准正态分布的密度曲线极为接近. 可以证明

$$\lim_{n} p_n(x) = \varphi(x) \quad (\text{一切 } x)$$

这里 $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} \exp\left\{-\dfrac{1}{2}x^2\right\}$ 是标准正态分布密度. 证明是不难的，请读者自己完成(提示：用斯特林公式).

t 分布的分位数表已造好，见书末附表 2.

据上所述，若 $X_1, \cdots, X_n$ 是来自 $N(\mu, \sigma^2)$ 的样本，对给定的置信水平 γ，查 t 分布表可找到 λ 满足(3.9)，从而得到 μ 的置信区间(3.10).

例 3.2 用某仪器间接测量温度，重复测量 5 次，得(单位：℃)：

1 250，1 265，1 245，1 260，1 275

问：温度的真值在什么范围？

我们很容易把这个问题化成数学问题. 用 μ 表示温度的真值，X 表示测量值. X 通常服从正态分布，$EX = \mu$(假定仪器无系统偏差). 上述 5 个数据可看成是来自 X 的样本. 问题是找出 μ 的置信区间(方差未知).

利用上面的理论，μ 的置信区间是(3.10). 现在 $n=5$，$\overline{X}=1\ 259$，$S^2 = \dfrac{1}{n-1}\sum_{1}^{n}(X_i - \overline{X})^2 = \dfrac{570}{4}$. 为了找 λ，查附表 2(注意自由度

$=n-1=4$),对于 $\gamma=0.95=1-0.05$,查得 $\lambda=2.776$,于是 μ 的置信区间是[1 244.2,1 273.8](置信水平为 0.95).

以上对指数分布和正态分布的参数讨论了如何利用枢轴量方法寻找置信区间. 对许多其他比较简单的分布,也可类似地得到参数的置信区间,这里不一一介绍了.

枢轴量方法有两个明显的缺点:① 怎样寻找枢轴量? 没有统一的方法;② 利用枢轴量方法得到的置信区间有何优良性质? 没有明确的结论.

现在介绍寻求置信区间的第二个方法——统计量方法. 设 $X_1,\cdots,X_n$ 是来自分布函数为 $F(x,\theta)$ $(\theta\in\Theta)$ 的总体的样本,Θ 是任意非空集合,$g(\theta)$ 是 Θ 上的实值函数. 设 $\varphi(X_1,\cdots,X_n)$ 是任何广义实值统计量,令

$$G(c,\theta)=P_\theta(\varphi(X_1,\cdots,X_n)\geqslant c) \tag{3.12}$$

$$H(c,\theta)=P_\theta(\varphi(X_1,\cdots,X_n)>c), \quad (-\infty\leqslant c\leqslant\infty) \tag{3.13}$$

这里 P_θ 是与参数 θ 相应的概率. 易知 $G(c,\theta)$ 是 c 的左连续减函数. 给定 $0<\gamma<1$,令

$$g_L(c)=\inf\{g(\theta):\theta\in\Theta \text{ 且 } G(c,\theta)>1-\gamma\} \text{①} \tag{3.14}$$

$$g_U(c)=\sup\{g(\theta):\theta\in\Theta \text{ 且 } H(c,\theta)<\gamma\} \tag{3.15}$$

我们有下列重要结论:

定理 3.5 (1) $g_L(\varphi(X_1,\cdots,X_n))$ 是 $g(\theta)$ 的置信水平为 γ 的置信下限,即

$$P_\theta(g(\theta)\geqslant g_L(\varphi(X_1,\cdots,X_n)))\geqslant\gamma \quad (\theta\in\Theta) \tag{3.16}$$

(2) $g_U(\varphi(X_1,\cdots,X_n))$ 是 $g(\theta)$ 的置信水平为 γ 的置信上限,即

$$P_\theta(g(\theta)\leqslant g_U(\varphi(X_1,\cdots,X_n)))\geqslant\gamma \quad (\theta\in\Theta) \tag{3.17}$$

(3) 设 $g_1=\min(g_L(\varphi(X_1,\cdots,X_n)),\varphi_U(\varphi(X_1,\cdots,X_n)))$,

---

① 当不存在 θ 满足 $G(c,\theta)>1-\gamma$ 时,令 $g_L(c)=\sup\{g(\theta):\theta\in\Theta\}$;当不存在 θ 满足 $H(c,\theta)<\gamma$ 时,令 $g_U(c)=\inf\{g(\theta):\theta\in\Theta\}$

$g_2=\max(g_L(\varphi(X_1,\cdots,X_n)),\varphi_U(\varphi(X_1,\cdots,X_n)))$，则$[g_1,g_2]$是$g(\theta)$的置信水平为 $2\gamma-1$ 的置信区间(当 $0.5<\gamma<1$ 时).

证明:(3)显然可从结论(1)和(2)直接推出. 下面证结论(1)成立.

固定 $\theta\in\Theta$,我们来证明

$$P_\theta(g(\theta))<g_L(\varphi(X_1,\cdots,X_n)))\leqslant 1-\gamma \tag{3.18}$$

为简单计,记 $Z=\varphi(X_1,\cdots,X_n)$,则 $G(u,\theta)=P_\theta(Z\geqslant u)$是 u 的减函数. 令

$$c_0=\inf\{c:-\infty\leqslant c\leqslant\infty,G(c,\theta)\leqslant 1-\gamma\}$$

分两种情形进行讨论.

① $G(c_0,\theta)\leqslant 1-\gamma$

此时 $P_\theta(g(\theta)<g_L(Z))\leqslant P_\theta(G(Z,\theta)\leqslant 1-\gamma)=P_\theta(Z\geqslant c_0)=G(c_0,\theta)\leqslant 1-\gamma$. 故(3.18)成立.

② $G(c_0,\theta)>1-\gamma$

此时 $P_\theta(g(\theta)<g_L(Z))\leqslant P_\theta(G(Z,\theta)\leqslant 1-\gamma)=P_\theta(Z>c_0)\overset{\text{记}}{=\!=}A$.

分三种情形考察. (i) $c_0=\infty$,此时 $A=0$;(ii) $c_0=-\infty$,此时 $A=\lim\limits_n P_\theta(Z\geqslant -n)=\lim\limits_n G(-n,\theta)\leqslant 1-\gamma$;(iii) $-\infty<c_0<\infty$,此时 $A=\lim\limits_n P_\theta\left(Z\geqslant c+\frac{1}{n}\right)=\lim\limits_n G\left(c_0+\frac{1}{n},\theta\right)\leqslant 1-\gamma$. 总之 $A\leqslant 1-\gamma$,故(3.18)成立. 从(3.18)直接推知(3.16)成立. 用同样的方法可证(3.17)成立. 定理 3.5 证毕.

从一个统计量 $\varphi=\varphi(X_1,\cdots,X_n)$出发,利用定理 3.5 得到 $g(\theta)$的置信下(上)限及置信区间的方法叫做统计量方法. 实际应用此方法时,通常取 $g(\theta)$的一个估计量或估计量的一个增函数作为统计量 $\varphi(X_1,\cdots,X_n)$,而且要使得 $G(c,\theta)$和 $H(c,\theta)$,$g_L(c)$,$g_U(c)$都比较好计算,才便于获得具体的置信限或置信区间. 这些都需要具体问题具体分析.

例 3.3 设 X 服从伯努利分布,即

$$P(X=1)=p=1-P(X=0),$$

其中 p 是未知参数，$0\leqslant p\leqslant 1$. 问：如何从 X 的样本 $X_1,\cdots,X_n$ 找出 p 的置信下限？（置信水平是 γ）.

易知，p 的矩估计 $\hat{p}=\sum_1^n X_i/n$，取统计量 $\varphi(X_1,\cdots,X_n)=\sum_1^n X_i$. 令

$$G(k,p)=P(\varphi(X_1,\cdots,X_n)\geqslant k)$$

当 $k=0,1,\cdots,n$ 时 $G(k,p)=P\left(\sum_1^n X_i\geqslant k\right)=\sum_{i=k}^n \mathrm{C}_n^i p^i(1-p)^{n-i}$.
多次使用分部积分公式得

$$G(k,p)=\frac{n!}{(k-1)!(n-k)!}\int_0^p x^{k-1}(1-x)^{n-k}\mathrm{d}x \quad (k\geqslant 1)$$

由此可见，$1\leqslant k\leqslant n$ 时 $G(k,p)$ 是 p 的严格增连续函数，设 $\underline{p}(k)$ 是方程 $G(k,p)=1-\gamma$ 的唯一根，则 $k\geqslant 1$ 时从(3.14)知

$$\begin{aligned}p_L(k)&=\inf\{p:0\leqslant p\leqslant 1,G(k,p)>1-\gamma\}\\&=\underline{p}(k)\end{aligned}$$

另一方面，$G(0,p)=1$，从(3.14)知 $p_L(0)=0$. 记 $\underline{p}(0)=0$ 则 p 的 γ 水平置信下限为 $\underline{p}\left(\sum_{i=1}^n X_i\right)$. 易知 $G(1,p)=1-(1-p)^n$，$G(n,p)=p^n$，故 $\underline{p}(1)=1-\gamma^{\frac{1}{n}}$，$\underline{p}(n)=(1-\gamma)^{\frac{1}{n}}$. 当 $1<k<n$ 时 $\underline{p}(k)$ 无显式表达，下一章将给出计算公式.

我们特别指出，$\underline{p}(n)=(1-\gamma)^{\frac{1}{n}}$ 是工程上应用颇广的重要公式. 例如，为了估计某种炮弹的发射成功率 p，进行了 20 次试验，结果每次都成功，则 p 的 0.80 水平置信下限 $p_L=(0.2)^{\frac{1}{20}}=0.9227$. $(1-\gamma)^{\frac{1}{n}}$ 乃是无失效情形下成功率的 γ 水平置信下限.

自然要问：是否有 $\psi(X_1,\cdots,X_n)$ 满足

$$P_p(p\geqslant\psi(X_1,\cdots,X_n))\equiv\gamma \quad (\text{一切 } p\in(0,1)) \qquad (3.19)$$

答案是，这样的 $\psi(X_1,\cdots,X_n)$ 不存在. 用反证法. 假设有这样的 ψ. 记 $a=\psi(1,\cdots,1)$（自变量全是 1）. 分两种情况讨论.

① $a<1$，此时当 $p\in(a,1)$ 时 $\gamma\equiv P_p(p\geqslant\psi(X_1,\cdots,X_n))\geqslant P_p(p\geqslant\psi(X_1,\cdots,X_n),X_1=\cdots=X_n=1)=P_p(X_1=\cdots=X_n=1)=p^n$. 令 $p\to1$ 得 $\gamma\geqslant1$，矛盾！② $a\geqslant1$，此时对一切 $p\in(0,1)$ 有 $1-\gamma\equiv P_p(p<\psi(X_1,\cdots,X_n))\geqslant P_p(p<\psi(X_1,\cdots,X_n),X_1=\cdots=X_n=1)=p^n$. 令 $p\to1$ 得 $1-\gamma\geqslant1$. 这又与 $\gamma\in(0,1)$ 相矛盾. 总之，不可能有 $\psi(X_1,\cdots,X_n)$ 满足(3.17).

用同样的方法也可以证明，对任何 $\gamma\in(0,1)$ 不存在 $\psi_1(X_1,\cdots,X_n)\leqslant\psi_2(X_1,\cdots,X_n)$ 满足

$$P_p(\psi_1(X_1,\cdots,X_n)\leqslant p\leqslant\psi_2(X_1,\cdots,X_n))\equiv\gamma$$

(对一切 $p\in(0,1)$).

但是，可以证明

$$\inf_p P_p\left(p\geqslant\underline{p}\left(\sum_1^n X_i\right)\right)=\gamma \tag{3.20}$$

这里 $\underline{p}(k)$ 的定义见前，即 $\underline{p}(0)=0$，$k\geqslant1$ 时，$\underline{p}(k)$ 是方程 $G(k,p)=1-\gamma$ 的唯一根.

我们可证明下列一般性结论：

定理 3.6 设 $G(c,\theta)$ 和 $g_L(c)$ 分别由(3.12)和(3.14)给出. 若对任何 $\varepsilon>0$，存在样本值 $(x_1^0,\cdots,x_n^0)$ 及 $\theta^*\in\Theta$ 满足下列条件：

(1) $g_L(\varphi(x_1^0,\cdots,x_n^0))=g(\theta^*)$

(2) $G(\varphi(x_1^0,\cdots,x_n^0),\theta^*)\geqslant1-\gamma-\varepsilon$

(3) 有一列 $\theta_1,\theta_2,\cdots$ 使得 $g(\theta_k)<g(\theta^*)$ 且 $\lim\limits_k G(\varphi(x_1^0,\cdots,x_n^0),\theta_k)=G(\varphi(x_1^0,\cdots,x_n^0),\theta^*)$，则

$$\inf_{\theta\in\Theta}P_\theta(g(\theta)\geqslant g_L(\varphi(X_1,\cdots,X_n)))=\gamma \tag{3.21}$$

证明：记 $Z=\varphi(X_1,\cdots,X_n)$，$z_0=\varphi(x_1^0,\cdots,x_n^0)$. 从(3.16)知

$$\begin{aligned}\gamma&\leqslant\inf_{\theta\in\Theta}P_\theta(g(\theta)\geqslant g_L(Z))\leqslant P_{\theta_k}(g(\theta_k)\geqslant g_L(Z))\\&\leqslant P_{\theta_k}(g(\theta^*)>g_L(Z))=P_{\theta_k}(g_L(z_0)>g_L(Z))\\&=1-P_{\theta_k}(g_L(Z)\geqslant g_L(z_0))\leqslant1-P_{\theta_k}(Z\geqslant z_0)=1-G(z_0,\theta_k).\end{aligned}$$

令 $k\to\infty$ 得

$$\gamma\leqslant\inf_{\theta\in\Theta}P_\theta(g(\theta)\geqslant g_L(Z))\leqslant1-G(z_0,\theta^*)\leqslant\gamma+\varepsilon.$$

再令 $\varepsilon\to 0$ 知(3.21)成立. 证毕.

对于置信上限 $g_U(\varphi(X_1,\cdots,X_n))$有类似的定理,从略.

从定理 3.6 容易推知(3.20)成立. (取 $x_1^0=\cdots=x_n^0=1, g(\theta)=\theta, \theta=p$, $\theta^*=p^*=(1-\gamma)^{\frac{1}{n}}, \theta_k=p_k \overset{\mathrm{d}}{=} p^*\left(1-\frac{1}{k}\right), k=1,2,\cdots$,即知).

从定理 3.5 和定理 3.6 的证明过程知道,我们并没有利用 $X_1,\cdots,X_n$ 是"简单随机样本"的性质,只要$(X_1,\cdots,X_n)$是随机向量,其概率分布依赖于参数 θ 即可. 因而统计量方法应用极广. 既然从任何一个统计量 $\varphi(X_1,\cdots,X_n)$出发都可得到 $g(\theta)$的置信限(下限或上限),那么自然要问:这样得到的置信限有何优良性?

为了表述优良性,先下一定义.

定义 3.4 设 $\varphi(x_1,\cdots,x_n)$和 $\psi(x_1,\cdots,x_n)$是两个函数,称 ψ 对 φ 是保序的,若对任何$(x_1,\cdots,x_n)$和$(x'_1,\cdots,x'_n)$,只要 $\varphi(x_1,\cdots,x_n)\leqslant\varphi(x'_1,\cdots,x'_n)$就一定成立 $\psi(x_1,\cdots,x_n)\leqslant\psi(x'_1,\cdots,x'_n)$.

我们可以证明下列

定理 3.7 设 $\varphi(X_1,\cdots,X_n)$是任何统计量,$g(\theta)$是 θ 的函数,$g_L(\varphi(X_1,\cdots,X_n))$和 $g_U(\varphi(X_1,\cdots,X_n))$分别由(3.14)和(3.15)确定. $\psi_1(X_1,\cdots,X_n)$和 $\psi_2(X_1,\cdots,X_n)$分别是 $g(\theta)$的 γ 水平置信下限和置信上限且二者都对 φ 是保序的,则

$$\psi_1(X_1,\cdots,X_n)\leqslant g_L(\varphi(X_1,\cdots,X_n))$$
$$\psi_2(X_1,\cdots,X_n)\geqslant g_U(\varphi(X_1,\cdots,X_n))$$

换句话说,由统计量 φ 导出的置信下限 $g_L(\varphi)$是所有对 φ 保序的置信下限中最大的,由 φ 导出的置信上限 $g_U(\varphi)$是所有对 φ 保序的置信上限中最小的. (注意,对给定的置信水平而言,上限越小越好,下限越大越好.)

我们不叙述定理 3.7 的证明了,读者如有兴趣,可参看[16]的第六章.

顺便说一句,统计量方法有一个等价形式——样本空间排序法,后者有时更便于应用(见[16]的第六章).

作为本节的结尾,我们指出下列两点:

(1) 我们未对置信区间的优良性进行深入讨论,迄今尚缺乏统一的令人满意的理论. 对一些特殊情形是有严密的数学理论的,参看文献[2],[12]. 当然,我们应该找宽度尽可能小的置信

区间.

(2) 置信区间的宽度一般是随机变量. 对于 $N(\mu,\sigma^2)$来说，当 σ 未知时，置信区间(3.10)的宽度是

$$l = 2\lambda\sqrt{\frac{1}{n(n-1)}\sum_{1}^{n}(X_i - \overline{X})^2}$$

l 取值可能很大，不能控制. 自然要问：是否有固定宽度的置信区间呢？可以证明，对于 $N(\mu,\sigma^2)$，只要 σ 未知，不管样本量 n 多么大，μ 不存在固定宽度的置信区间. 但可采用两阶段方法(样本量是随机变量)得到固定宽度的置信区间，具体理由见第六章.

§4 分布函数与密度函数的估计

分布函数的估计　设 X 的分布函数 $F(x)$未知，如何从样本 $X_1,\cdots,X_n$ 估计 $F(x)$的值(x 给定)？这是很重要的问题. 掌握随机变量的特性就是要弄清它的分布函数，许多重要的实际问题要求估计 $F(x)$. 例如在产品或生物体的寿命分析中，人们常常关心寿命 X 大于 x 的概率 $S(x)$，显然 $S(x)=1-F(x)$. $S(x)$叫做可靠性函数或生存函数.

如果 $F(x)$的函数类型已知，只是含有未知的参数，则常用前几节的办法估计未知参数，从而弄清 $F(x)$是什么. 若不知道 $F(x)$的分布类型，怎么办？只好用非参数方法.

给定 x 后，记 ν_n 为 $X_1,\cdots,X_n$ 中不超过 x 的个数，自然用频率$F_n(x)\xlongequal{\mathrm{d}}\frac{1}{n}\nu_n$ 去估计概率 $F(x)=P(X\leqslant x)$.

定义 4.1　称 x 的函数 $F_n(x)$为 X 的经验分布函数.

注意，

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{(-\infty,x]}(X_i)$$

这里 I_A 表示 A 的示性函数. 根据强大数律知

$$P(\lim_n F_n(x)=F(x))=1 \qquad (\text{一切 } x)$$

这是经验分布函数逼近理论分布函数的理论根据. $F_n(x)$的表达式可以写得更清楚些. 将 $X_1,\cdots,X_n$ 按值从小到大排列,得到 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 其中 $X_{(i)}$ 叫做第 i 个次序统计量($i=1,2,\cdots,n$). 易知

$$F_n(x)=\begin{cases}0, & x<X_{(1)}\\ \dfrac{k}{n}, & x\in[X_{(k)},X_{(k+1)})\ (k=1,2,\cdots,n-1)\\ 1, & x\geqslant X_{(n)}\end{cases}$$

我们有下列更强的结论:

定理 4.1(Glivenko - Cantelli) 设

$$D_n=\sup_x|F_n(x)-F(x)|$$

则

$$P(\lim_n D_n=0)=1$$

证明:设 γ 是任一正整数,$k=1,2,\cdots,\gamma$. 令

$$x_{\gamma,k}=\inf\left\{x:F(x)\geqslant\frac{k}{\gamma}\right\}\qquad(\inf\varnothing\overset{\text{d}}{=}\infty)$$

设 $x\in[x_{\gamma,k},x_{\gamma,k+1})$,则

$$F_n(x_{\gamma,k})\leqslant F_n(x)\leqslant F_n(x_{\gamma,k+1}-0)$$

$$F(x_{\gamma,k})\leqslant F(x)\leqslant F(x_{\gamma,k+1}-0)$$

于是

$$F_n(x_{\gamma,k})-F(x_{\gamma,k+1}-0)\leqslant F_n(x)-F(x)$$
$$\leqslant F_n(x_{\gamma,k+1}-0)-F(x_{\gamma,k})$$

由于

$$F(x_{\gamma,k}-0)\leqslant\frac{k}{\gamma}\leqslant F(x_{\gamma,k})\leqslant F(x_{\gamma,k+1}-0)$$
$$\leqslant\frac{k+1}{\gamma}\leqslant F(x_{\gamma,k+1})$$

故

$$0\leqslant F(x_{\gamma,k+1}-0)-F(x_{\gamma,k})\leqslant\frac{1}{\gamma}$$

从而

$$F_n(x)-F(x)\leqslant F_n(x_{\gamma,k+1}-0)-F(x_{\gamma,k+1}-0)+\frac{1}{\gamma}$$

$$F_n(x)-F(x)\geqslant F_n(x_{\gamma,k})-F(x_{\gamma,k})-\frac{1}{\gamma}$$

于是

$$|F_n(x)-F(x)|\leqslant |F_n(x_{\gamma,k+1}-0)-F(x_{\gamma,k+1}-0)| + |F_n(x_{\gamma,k})-F(x_{\gamma,k})|+\frac{1}{\gamma}$$

其中 $x\in[x_{\gamma,k},x_{\gamma,k+1})$ $(k=1,2,\cdots,\gamma-1)$.

当 $x<x_{\gamma,1}$ 时，

$$F_n(x)-F(x)\leqslant F_n(x_{\gamma,1}-0)\leqslant|F_n(x_{\gamma,1}-0)-F(x_{\gamma,1}-0)|+\frac{1}{\gamma},$$

$$F(x)-F_n(x)\leqslant F(x_{\gamma,1}-0)\leqslant\frac{1}{\gamma},$$

从而

$$|F_n(x)-F(x)|\leqslant |F_n(x_{\gamma,1}-0)-F(x_{\gamma,1}-0)|+\frac{1}{\gamma}$$

当 $x\geqslant x_{\gamma,\gamma}$ 时，$F(x)=1$，故

$$|F_n(x)-F(x)|=F(x)-F_n(x)\leqslant F(x_{\gamma,\gamma})-F_n(x_{\gamma,\gamma}) = |F_n(x_{\gamma,\gamma})-F(x_{\gamma,\gamma})|$$

（注意，我们规定 $F(\infty)=F_n(\infty)=1$).

总之有

$$\sup_x|F_n(x)-F(x)|\leqslant\sum_{k=1}^{\gamma}|F_n(x_{\gamma,k}-0)-F(x_{\gamma,k}-0)| + \sum_{k=1}^{\gamma}|F_n(x_{\gamma,k})-F(x_{\gamma,k})|+\frac{1}{\gamma}$$

令

$$A_{\gamma,k}=\{\lim_n F_n(x_{\gamma,k}-0)=F(x_{\gamma,k}-0)$$

且 $\qquad \lim\limits_n F_n(x_{\gamma,k})=F(x_{\gamma,k})\}$ $\quad(k=1,2,\cdots,\gamma)$

由于

$$F_n(x)=\frac{1}{n}\sum_1^n I_{(-\infty,x]}(X_i),F_n(x-0)=\frac{1}{n}\sum_1^n I_{(-\infty,x)}(X_i)$$

从强大数律知

$$P(A_{\gamma,k})=1 \quad (\gamma\geqslant 1,k=1,\cdots,\gamma)$$

令 $A=\bigcap_{r=1}^{\infty}\bigcap_{k=1}^{r}A_{\gamma,k}$，易知 $P(A)=1$. 但是

$$A\subset\{\lim_{n}D_n=0\}$$

所以

$$P(\lim_{n}D_n=0)=1$$

证毕.

经过现代的深入研究，还可以证明：对任何 $\varepsilon>0$，存在常数 C（与 n 无关），使得

$$P(D_n>\varepsilon)\leqslant Ce^{-n\varepsilon^2} \quad (n\geqslant 1)$$

在实际工作中有时需要估计分布函数的分位数. 设 x_p 是 $F(x)$ 的 p 分位数（$0<p<1$），即

$$F(x_p)\geqslant p\geqslant F(x_p-0)$$

（即 $P(X\leqslant x_p)\geqslant p\geqslant P(X<x_p)$）

若 $X_{(1)},\cdots,X_{(n)}$ 是样本 $X_1,\cdots,X_n$ 的次序统计量，$\gamma=[np]+1$，我们可用第 r 个次序统计量 $X_{(r)}$ 作为 x_p 的估计量. 可以证明，只要方程"$F(x)=p$"至多一个根（此时 x_p 唯一），则样本量无限增大时 $X_{(r)}$ 就是 x_p 的强相合估计.

例 4.1 某食品厂用自动装罐机生产额定净重为 345 g 的午餐肉罐头，由于随机性，每个罐头的净重都有些差别. 现从生产线上随机抽取 10 个罐头，秤其净重，得下列数据（单位：g）：

344 336 345 342 340 338

344 343 344 343

试求该生产线上生产出的罐头的净重的分布函数，并估计其中位数.

解 我们用经验分布函数 $F_{10}(x)$ 作为分布函数的估计. 将样本值从小到大排列：336，338，340，342，343，343，344，344，344，345，于是可得经验分布函数 $F_{10}(x)$ 如下：

$$F_{10}(x)=\begin{cases}0, & x<336\\ \dfrac{1}{10}, & 336\leqslant x<338\\ \dfrac{2}{10}, & 338\leqslant x<340\\ \dfrac{3}{10}, & 340\leqslant x<342\\ \dfrac{4}{10}, & 342\leqslant x<343\\ \dfrac{6}{10}, & 343\leqslant x<344\\ \dfrac{9}{10}, & 344\leqslant x<345\\ 1, & x\geqslant 345\end{cases}$$

这就是分布函数的近似值. 注意 $p=\dfrac{1}{2}$时,$[pn]+1=6$. 故可用 $x_{(6)}$作为中位数 $x_{\frac{1}{2}}$的估计. 即罐头净重的中位数约为 343 g.

如果随机变量 X 有密度函数 $p(x)$,则应研究密度函数 $p(x)$如何估计. 因为密度函数更能直观地刻画出概率分布的特性(如对称性、峰值等等),如何估计的问题显得很重要. 特别是多维随机向量情形,分布函数用处较小,又不便于处理,估计密度函数意义更大,在图像识别及多元判决中要用到. 这里仅讨论一维随机变量的密度估计问题. 估计方法有很多种,这里首先介绍历史悠久,现在仍在广泛使用的直方图法,然后介绍近 40 年发展起来的核估计法和最近邻估计法. 下面考虑的随机变量均是连续型的.

(一) 直方图法

设 $X_1,\cdots,X_n$ 是来自密度为 $f(x)$的总体的样本,用 $R_n(a,b)$表示诸 X_i 中落在区间$(a,b]$的个数. 显然积分 $\int_a^b f(x)\mathrm{d}x$ 可用频率$\dfrac{1}{n}R_n(a,b)$来估计. 由微分学中值定理知存在 $x_0\in(a,b]$

使得

$$f(x_0)=\frac{1}{b-a}\int_a^b f(x)\mathrm{d}x$$

故可用

$$f_n(x_0)=\frac{1}{n(b-a)}R_n(a,b)$$

作为 $f(x_0)$ 的估计值. 如果 $f(x)$ 在 $(a,b]$ 上连续, 又 $b-a$ 很小, 则可用 $f_n(x_0)$ 作为 $f(x)$ 的近似值(一切 $x\in(a,b]$).

更确切地叙述如下: 设 $\{t_i: i=0,1,\cdots,m\}$ 是 $m+1$ 个数: $-\infty<t_0<t_1<\cdots<t_m<\infty$. 通常假定 $t_{i+1}-t_i\equiv h>0(i=0,1,\cdots,m-1)$. 令

$$f_n(x)=\begin{cases}\dfrac{1}{nh}R_n(t_i,t_{i+1}), & \text{当 } x\in(t_i,t_{i+1}],i=0,1,\cdots,m-1\\ 0, & \text{当 } x\leqslant t_0 \text{ 或 } x>t_m\end{cases} \tag{4.1}$$

用 $f_n(x)$ 作为 $f(x)$ 的估计. 这就是直方图估计法.

实际使用时, 将样本值 $X_1,\cdots,X_n$ 从小到大排列, 得 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$, 取 a 为比 $X_{(1)}$ 略小的数, b 为比 $X_{(n)}$ 略大的数. 将区间 $(a,b]$ m 等分. 分点

$$t_i=a+i\frac{b-a}{m}\quad(i=0,1,\cdots,m)$$

设样本落入 $(t_i,t_{i+1}]$ 之频数为 ν_i, 在数轴上作以 $[t_i,t_{i+1}]$ 为底, 以 $\nu_i/(nh)$ 为高的长方形(这里 $h=t_{i+1}-t_i=(b-a)/m$), 这一列矩形叫做直方图. 注意

$$\text{第 } i \text{ 个长方形的面积}=\frac{\nu_i}{n}\doteq\int_{t_i}^{t_{i+1}}f(y)\mathrm{d}y$$

在直方图法中 m 的大小没有硬性规定, 当样本量 n 小时, m

也应小些；n 大时，m 则大些（比如，$n=100$ 时 m 可取 12）.①

我们详细叙述下列例子，读者可从中掌握直方图法.

例 4.2 某炼钢厂生产了一种钢 25Mn Si，由于各种偶然因素的影响，各炉钢的含 Si 量是有差异的，因而应把含 Si 量 X 看成一个随机变量，现在看看它的概率密度函数是怎样的？

为了确定密度函数，记录了 120 炉正常生产的 25Mn Si 钢的含 Si 量的数据（百分数）.

0.86	0.83	0.77	0.81	0.81	0.80
0.79	0.82	0.82	0.81	0.81	0.87
0.82	0.78	0.80	0.81	0.87	0.81
0.77	0.78	0.77	0.78	0.77	0.77
0.77	0.71	0.95	0.78	0.81	0.79
0.80	0.77	0.76	0.82	0.80	0.82
0.84	0.79	0.90	0.82	0.79	0.82
0.79	0.86	0.76	0.78	0.83	0.75
0.82	0.78	0.73	0.83	0.81	0.81
0.83	0.89	0.81	0.86	0.82	0.82
0.78	0.84	0.84	0.84	0.81	0.81
0.74	0.78	0.78	0.80	0.74	0.78
0.75	0.79	0.85	0.75	0.74	0.71
0.88	0.82	0.76	0.85	0.73	0.78

---

① 有人建议采用下列公式：

$$m \doteq 1 + 3.322\log_{10} n$$

也可按下表选择 m

n	m
<50	5～6
50～100	6～10
100～250	7～12
>250	10～20

0.81	0.79	0.77	0.78	0.81	0.87
0.83	0.65	0.64	0.78	0.75	0.82
0.80	0.80	0.77	0.81	0.75	0.83
0.90	0.80	0.85	0.81	0.77	0.78
0.82	0.84	0.85	0.84	0.82	0.85
0.84	0.82	0.85	0.84	0.78	0.78

下面对这 120 个数据进行分组：

(1) 找出它们的最小值为 0.64,最大值为 0.95.

(2) 取起点 $a=0.635$,终点 $b=0.955$,共分 $m=16$ 组,组距$h=0.02$.

(3) 分组及频数如下：

分组	频数 ν_i
0.635～0.655	2
0.655～0.675	0
0.675～0.695	0
0.695～0.715	2
0.715～0.735	2
0.735～0.755	8
0.755～0.775	13
0.775～0.795	23
0.795～0.815	24
0.815～0.835	21
0.835～0.855	14
0.855～0.875	6
0.875～0.895	2
0.895～0.915	2
0.915～0.935	0
0.935～0.955	1

以上用实例介绍了如何分组,下面根据分组情况及其频数来作直方图. 记 $f_i=\nu_i/n(i=0,1,\cdots,m-1)$. 设横轴 x 表示含 Si

量，这里 $a=t_0=0.635$，$t_1=0.655$，…，$t_{15}=0.935$，$t_{16}=b=0.955$，$t_{i+1}-t_i=0.02$. 在区间$[t_i,t_{i+1}]$($0\leqslant i\leqslant 15$)上作高度为 f_i/h 的长方形，这 16 个并立的长方形是本例的直方图. 注意，第 i 个长方形的高度是

$$\frac{f_i}{h}=\nu_i\cdot\frac{1}{120}\cdot\frac{1}{0.02}=\frac{\nu_i}{2.4}$$

为方便起见，取纵坐标的单位是$\frac{1}{nh}=\frac{1}{2.4}$，则直方图中第 i 个长方形的高度正好是 ν_i 个单位，见图 1.

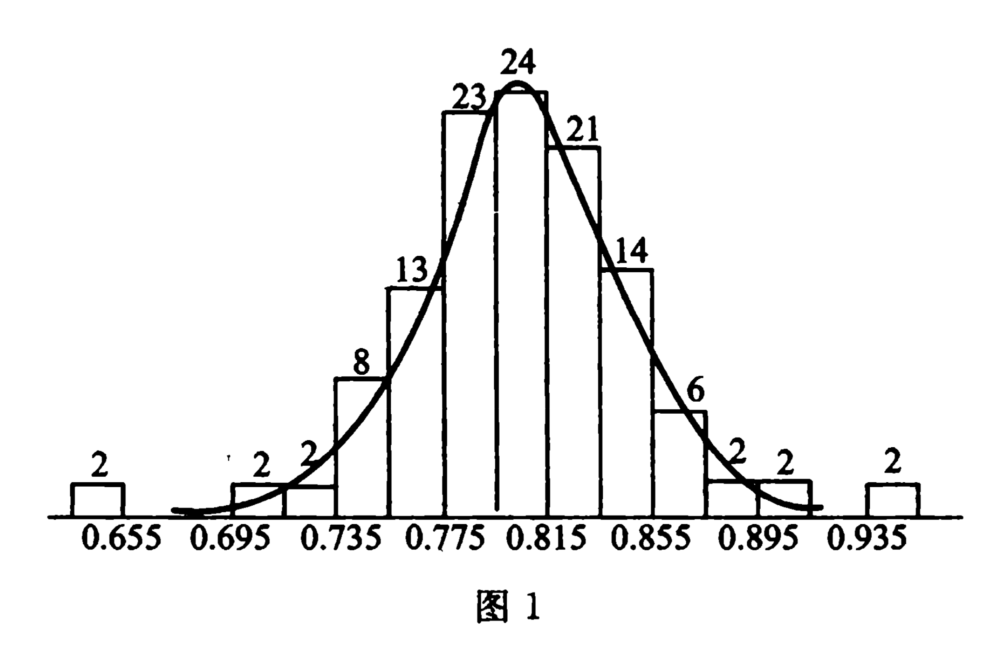


图 1

有了直方图，可以看出 X 的密度函数曲线大体是图中曲线的位置. 从图上看，这条曲线很像是正态分布密度的曲线. 怎样根据数据判断 X 是否服从正态分布呢？解决这个问题的办法是有的，请看第三章.

直方图估计的理论基础是什么呢？我们可以证明，只要间隔 h 足够小，样本量 n 充分大，则$\sup_x|f_n(x)-f(x)|$可任意小，这里 $f_n(x)$由(4.1)确定. 为了确切说明这一点，设$\{t_k^{(n)}\}$是直线上的点列：$\cdots<t_k^{(n)}<t_{k+1}^{(n)}<\cdots$，这里 $t_{k+1}^{(n)}-t_k^{(n)}\equiv h_n>0$(一切整数 k). 令

$$f_n(x)=\frac{1}{nh_n}R_n(t_k^{(n)},t_{k+1}^{(n)})\quad(\text{当 } x\in(t_k^{(n)},t_{k+1}^{(n)}])$$

这里

$$R_n(t_k^{(n)},t_{k+1}^{(n)}) = \sum_{i=1}^{n} I_{(t_k^{(n)},t_{k+1}^{(n)}]}(X_i)$$

定理 4.2 设 $f(\cdot)$在点 x 连续且$\lim\limits_n h_n=0$，$\lim\limits_n nh_n=\infty$，则对任何 $\varepsilon>0$ 有

$$\lim_n P(|f_n(x)-f(x)|\geqslant\varepsilon)=0 \tag{4.2}$$

证明：因为 $f(\cdot)$在点 x 连续，故有 $\delta>0$ 及 $K_0>0$，使对一切 $y\in[x-\delta,x+\delta]$有 $f(y)\leqslant K_0$.

设 n 充分大，使得 $h_n<\delta$. 取 k 使得 $x\in(t_k^{(n)},t_{k+1}^{(n)}]$. 易知

$$\begin{aligned} P(|f_n(x)-f(x)|\geqslant\varepsilon) &\leqslant \frac{1}{\varepsilon^2}E(f_n(x)-f(x))^2 \\ &\leqslant \frac{2}{\varepsilon^2}\{E(f_n(x)-Ef_n(x))^2 \\ &\quad +(Ef_n(x)-f(x))^2\} \end{aligned} \tag{4.3}$$

我们指出

$$f_n(x)-Ef_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}\alpha_n(X_i)$$

这里

$$\alpha_n(X_i)=I_{(t_k^{(n)},t_{k+1}^{(n)}]}(X_i)-E(I_{(t_k^{(n)},t_{k+1}^{(n)}]}(X_i))$$

$$E\alpha_n(X_i)=0$$

$$E(\alpha_n(X_i))^2\leqslant\int_{t_k^{(n)}}^{t_{k+1}^{(n)}}f(y)\mathrm{d}y\leqslant K_0h_n$$

于是

$$E[f_n(x)-Ef_n(x)]^2\leqslant\frac{nK_0h_n}{n^2h_n^2}=\frac{K_0}{nh_n}$$

另一方面，

$$Ef_n(x)=\frac{1}{h_n}\int_{t_k^{(n)}}^{t_{k+1}^{(n)}}f(y)\mathrm{d}y\to f(x)\quad(n\to\infty)$$

故从(4.3)推知(4.2)成立.

证毕.

经过更深入的分析，可以证明(见[13])：若密度函数 $f(x)$在

$(-\infty,\infty)$上一致连续，$\int_{-\infty}^{\infty}|x|^{\delta}f(x)\mathrm{d}x<\infty$（对某个$\delta>0$），又$\lim\limits_{n}h_n=0$，$h_n\geqslant\frac{1}{n}(\ln n)^2$，则

$$P(\lim_{n}\sup_{x}|f_n(x)-f(x)|=0)=1$$

这表明 $f_n(x)$有一致强相合性.

（二）核估计法

在引进一般的核估计之前，先讲一个特殊情形，以便读者理解核估计的思想. 设 X 的分布函数和密度函数分别是$F(x)$和$f(x)$. 若 $f(x)$连续，则 h 很小时有

$$\frac{F(x+h)-F(x-h)}{2h}\approx f(x)$$

而 $F(x)$可用经验分布函数 $F_n(x)$来估计，从而可用

$$\hat{f}_n(x)=\frac{1}{2h}[F_n(x+h)-F_n(x-h)]$$

来估计 $f(x)$，这叫做 $f(x)$的 Rosenblatt 估计，是 M. Rosenblatt 于 1956 年首先提出来的. 不难看出，

$$\hat{f}_n(x)=\frac{1}{2hn}\sum_{i=1}^{n}I_{(x-h,x+h]}(X_i)$$

令

$$K_0(x)=\begin{cases}\frac{1}{2}, & x\in[-1,1)\\ 0, & \text{其他}\end{cases}$$

则

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}K_0\left(\frac{x-X_i}{h}\right)$$

即 $\hat{f}_n(x)$可以通过一个“核函数”$K_0(x)$表达出来.

定义 4.2 设 $K(x)$是$(-\infty,\infty)$上的非负函数且$\int_{-\infty}^{\infty}K(x)\mathrm{d}x=1$，则称 $K(x)$是核函数. 此时称

$$\tilde{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\tag{4.4}$$

为 $f(x)$的核估计.

核函数 $K(x)$有很大的选择自由，类型很多，如

$$K_0(x)=\begin{cases}\frac{1}{2}, & -1 \leqslant x < 1 \\ 0, & 其他\end{cases}$$

$$K_1(x)=\begin{cases}1, & |x| \leqslant 1/2 \\ 0, & 其他\end{cases}$$

$$K_2(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},$$

$$K_3(x)=\frac{1}{\pi(1+x^2)}$$

$$K_4(x)=\frac{1}{2\pi}\left(\frac{\sin\frac{x}{2}}{\frac{x}{2}}\right)^2$$

可以证明，在一定条件下，核估计具有相合性.

定理 4.3 设核函数 $K(x)$ 满足条件：

$$\int_{-\infty}^{\infty}(K(x))^2\mathrm{d}x < \infty, \quad \lim_{|x|\to\infty}|x|K(x)=0$$

又密度函数 f 在点 x 连续，$h_n \to 0$ 且 $nh_n \to \infty (n \to \infty)$，则对一切 $\varepsilon > 0$ 有

$$\lim_n P(|\tilde{f}_n(x)-f(x)| \geqslant \varepsilon)=0$$

其中 $\tilde{f}_n(x)$ 是 $f(x)$ 的核估计：

$$\tilde{f}_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right)$$

证明：首先证明：

$$\lim_n E\tilde{f}_n(x) = f(x) \tag{4.5}$$

实际上，

$$E\tilde{f}_n(x)=\int_{-\infty}^{\infty}h_n^{-1}K\left(\frac{x-y}{h_n}\right)f(y)\mathrm{d}y$$

$$=\int_{-\infty}^{\infty}K(u)f(x-h_n u)\mathrm{d}u$$

$$|E\tilde{f}_n(x)-f(x)| = \left|\int_{-\infty}^{\infty}K(u)[f(x-h_n u)-f(x)]\mathrm{d}u\right|$$

$$\leqslant \sup_{|u|\leqslant \delta h_n^{-1}}|f(x-h_n u)-f(x)|$$

$$+\sup_{|u|>\delta h_n^{-1}}K(u)\cdot\int_{-\infty}^{\infty}f(x-h_n s)\mathrm{d}s$$

$$+ |f(x)| \cdot \int_{|u|>\delta h_n^{-1}} K(u)\mathrm{d}u$$

$$\leqslant \sup_{|y|\leqslant\delta} |f(x-y)-f(x)|$$

$$+\delta^{-1} \cdot \sup_{|u|>\delta h_n^{-1}} |uK(u)|$$

$$+ |f(x)| \int_{|u|>\delta h_n^{-1}} K(u)\mathrm{d}u \quad (\delta>0)$$

利用这个不等式,不难看出(4.5)成立.

现在来证明:

$$\lim_n \mathrm{Var}(\widetilde{f}_n(x)) = 0 \tag{4.6}$$

实际上,

$$\mathrm{Var}(\widetilde{f}_n(x)) = \frac{1}{n}\mathrm{Var}\left[\frac{1}{h_n}K\left(\frac{x-X_1}{h_n}\right)\right]$$

$$\leqslant \frac{1}{nh_n}\int_{-\infty}^{\infty} \frac{1}{h_n}\left[K\left(\frac{x-y}{h_n}\right)\right]^2 f(y)\mathrm{d}y$$

$$= \frac{1}{nh_n}\int_{-\infty}^{\infty} [K(u)]^2 f(x-h_n u)\mathrm{d}u$$

$$\leqslant \frac{1}{nh_n}\int_{-\infty}^{\infty} [K(u)]^2 \,|f(x-h_n u)-f(x)|\,\mathrm{d}u$$

$$+\frac{1}{nh_n}f(x)\int_{-\infty}^{\infty}[K(u)]^2\mathrm{d}u$$

$$\leqslant \frac{1}{nh_n}\sup_{|y|<\delta}|f(x-y)-f(x)| \cdot \int_{-\infty}^{\infty}[K(u)]^2\mathrm{d}u$$

$$+\frac{1}{nh_n}\int_{|u|>\delta h_n^{-1}}[K(u)]^2 f(x-h_n u)\mathrm{d}u$$

$$+\frac{2}{nh_n}f(x)\int_{-\infty}^{\infty}(K(u))^2\mathrm{d}u$$

$$\leqslant o(1)+\frac{1}{nh_n^2}\sup_{|u|>\delta h_n^{-1}}[K(u)]^2 \quad (n\to\infty)$$

$$\leqslant o(1)+\frac{1}{n\delta^2}\sup_{|u|>\delta h_n^{-1}}[uK(u)]^2 \quad (n\to\infty)$$

$$= o(1) \qquad (n\to\infty)$$

这就证明了(4.6)成立.

于是

$$P(|\tilde{f}_n(x)-f(x)|\geqslant\varepsilon)\leqslant\frac{1}{\varepsilon^2}E[\tilde{f}_n(x)-f(x)]^2$$

$$\leqslant\frac{2}{\varepsilon^2}\{\mathrm{Var}[\tilde{f}_n(x)]+[E\tilde{f}_n(x)-f(x)]^2\}$$

$$=o(1)\qquad(n\to\infty)$$

证毕.

经过更深入的分析,可以证明(见[13]):若密度函数 $f(x)$ 在 $(-\infty,\infty)$ 上一致连续,且

$$\lim_n h_n=0,\ \sum_{n=1}^{\infty}\exp\{-\gamma nh_n^2\}<\infty\quad(\text{一切 }\gamma>0)$$

又核函数是有界变差的,则

$$P(\lim_n\sup_x|\tilde{f}_n(x)-f(x)|=0)=1$$

只要核函数选得适当,核估计往往比直方图估计有较好的精度,我们不加证明地指出下列事实:设 $\hat{f}_n(x)$ 是前面介绍过的 Rosenblatt 估计(注意,这是一种特殊的核估计!),而 $f_n(x)$ 是直方图估计,即

$$f_n(x)=\frac{1}{nh_n}R_n(t_k^{(n)},t_{k+1}^{(n)}),\text{当 } x\in(t_k^{(n)},t_{k+1}^{(n)}],$$

则当 $h_n\to 0$ 且 $nh_n\to\infty(n\to\infty)$ 时有

$$E[\hat{f}_n(x)-f(x)]^2=O(n^{-\frac{4}{5}})$$

$$E[f_n(x)-f(x)]^2=O(n^{-\frac{2}{3}})$$

换句话说,Rosenblatt 估计比直方图估计有较小的渐近均方误差.

(三) 最近邻估计法

设 $f(x)$ 是 X 的分布密度,$X_1,\cdots,X_n$ 是样本. 1965 年 Loftsgarden 和 Quesenberry 提出下列最近邻估计法. 选定自然数 $K(n)$. 令

$$a_n(x)=\min\{t:R_n(x-t,x+t)\geqslant K(n)\}$$

$$f_n^*(x)=\frac{K(n)}{2na_n(x)}$$

其中

$$R_n(x-t,x+t)=\sum_{i=1}^{n}I_{(x-t,x+t]}(X_i)$$

是样本落入区间$(x-t,x+t]$的频数.

这个$f_n^*(x)$叫做$f(x)$的最近邻估计.

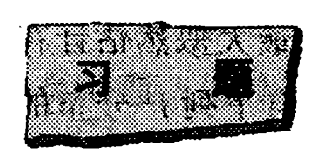
经过深入研究,可以证明(见[13])下列结论:若$f(x)$在$(-\infty,\infty)$上一致连续,且

$$\lim_n \frac{K(n)}{n}=0,\quad \lim_n \frac{K(n)}{\ln n}=\infty$$

则

$$P(\lim_n \sup_x |f_n^*(x)-f(x)|=0)=1$$

即$f_n^*(x)$是$f(x)$的一致强相合估计.



1. 设X服从几何分布

$$P(X=k)=p(1-p)^{k-1},\quad k=1,2,\cdots$$

$X_1,X_2,\cdots,X_n$是X的简单随机样本,试找出p的最大似然估计.

2. 设X的分布密度是

$$f(x)=\frac{1}{2\sigma}e^{-\frac{1}{\sigma}|x|}\qquad(\sigma>0)$$

$X_1,X_2,\cdots,X_n$是X的样本,试求σ的最大似然估计.

3. 设$X_1,X_2,\cdots,X_n$是来自$[\theta,\theta+1]$上均匀分布的样本,其中$-\infty<\theta<+\infty$. 试证明θ的最大似然估计不止一个. 你能求出θ的全部最大似然估计吗?你能找出θ的无偏估计吗?

4. 设随机变量X以均等机会按$N(0,1)$分布取值和按$N(\mu,\sigma^2)$分布取值(μ,σ^2未知,$-\infty<\mu<+\infty,\sigma^2>0$). 这时$X$的分布密度为这两个分布的密度的平均:

$$f(x;\mu,\sigma^2)=\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}+\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

设$X_1,X_2,\cdots,X_n$为此混合分布的简单随机样本,试证明μ和σ^2不存在最大似然估计.

5. 在买面包作早点的男、女消费者中,男性购买者的比例p未知,但知

道$\frac{1}{2}\leqslant p\leqslant\frac{2}{3}$. 设在 70 个购买者中发现 12 个是男性,58 个是女性,求 p 的最大似然估计.

如果对 p 无限制($0\leqslant p\leqslant 1$),求 p 的最大似然估计.

6. 设 $X_1,X_2,\cdots,X_n$ 是来自下列两参数指数分布的样本:

$$f(x;\theta_1,\theta_2)=\begin{cases}\frac{1}{\theta_2}e^{-\frac{1}{\theta_2}(x-\theta_1)}, & x\geqslant\theta_1\\ 0, & x<\theta_1\end{cases}$$

其中 $\theta_1\in(-\infty,+\infty)$,$\theta_2\in(0,+\infty)$,试求出 θ_1 和 θ_2 的最大似然估计.

7. 设 $X_1,X_2,\cdots,X_n$ 是来自下列 Weibull 分布的样本:

$$F(x;\alpha,\eta)=\begin{cases}1-e^{-\left(\frac{x}{\eta}\right)^\alpha}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

其中 $\alpha>0,\eta>0$,试证明 α,η 的最大似然估计存在.

8. 设 $X_1,X_2,\cdots,X_n$ 是来自下列 Γ 分布的样本:

$$f(x;\alpha,\beta)=\begin{cases}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\alpha>0,\beta>0$,试证明 α,β 的最大似然估计存在.

9. 设 $X_1,X_2,\cdots,X_n$ 是来自下列分布密度的样本:

$$f(x,\theta)=\begin{cases}\frac{\Gamma(\theta+1)}{\Gamma(\theta)\Gamma(1)}x^{\theta-1}, & 0\leqslant x\leqslant 1\\ 0, & \text{其他}\end{cases}$$

其中 $\theta\in(0,+\infty)$,试用矩法估计 θ.

10. 设 $X_1,X_2,\cdots,X_n$ 是来自分布密度为

$$f(x;c,\theta)=\frac{1}{2\theta}I_{[c-\theta,c+\theta]}(x)$$

的总体的样本,$-\infty<c<+\infty,\theta>0$,试用矩法估计 c 和 θ.

11. 设 $X_1,X_2,\cdots,X_n$ 为 $N(\mu,\sigma^2)$分布的样本,参数 μ,σ^2 未知($-\infty<\mu<+\infty,\sigma^2>0$),令

$$S_0^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

$$S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

证明对所有可能的 μ 和 σ^2 的值，作为 σ^2 的估计，S_0^2 的均方误差比 S_1^2 的均方误差都小，即 S_0^2 比 S_1^2 有效；但 S_1^2 是无偏估计而 S_0^2 不是；S_0^2 是最大似然估计.

12. 设 $X_1,X_2,\cdots,X_n$ 是来自下列的 $\Gamma(\alpha,\beta)$ 分布的样本：

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\alpha>0,\beta>0$ 未知，证明 $\left(\prod\limits_{i=1}^n X_i,\sum\limits_{i=1}^n X_i\right)$ 是 (α,β) 的充分统计量.

13. 设 $X_1,X_2,\cdots,X_n$ 是来自参数为 λ 的 Poisson 分布的样本，试证明 $\overline{X}$ 和 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^n(X_i-\overline{X})^2$ 都是 λ 的无偏估计，并问：是否还有其他的无偏估计？

14. 设 $X_1,X_2,\cdots,X_n$ 是来自下列二项分布的样本：

$$P(X=k)=C_m^k\theta^k(1-\theta)^{m-k},\quad k=0,1,\cdots,m$$

试求 θ^2 的无偏估计.

15. 设 $X_1,X_2,\cdots,X_n$ 是来自参数为 λ 的 Poisson 分布的样本，试求 λ^2 的无偏估计.

16. 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，试证明

$$\hat{\sigma}=\frac{1}{\sqrt{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sqrt{\sum_{i=1}^n(X_i-\overline{X})^2}$$

是 σ 的无偏估计.

17. 设 $X_1,X_2,\cdots,X_n$ 是来自如下密度的 Pareto 分布的样本：

$$f(x;\theta)=\begin{cases}\dfrac{\theta}{(1+x)^{\theta+1}}, & 0<x<+\infty\\ 0, & x\leqslant 0\end{cases}$$

其中 $\theta\in(1,+\infty)$，试找出 θ^{-1} 的无偏估计.

18. 设 $X_1,X_2,\cdots,X_n$ 是来自下列分布密度的总体的样本：

$$f(x;\theta)=\begin{cases}\dfrac{1}{\theta}e^{-\frac{1}{\theta}x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

其中 $\theta>0$，试找出 θ 的无偏估计的方差的尽可能大的下界.

19. 设 $X_1,X_2,\cdots,X_n$ 是来自如下密度的 Cauchy 分布的样本：

$$f(x;\theta)=\frac{1}{\pi[1+(x-\theta)^2]}$$

其中 $\theta\in(-\infty,+\infty)$，试找出 θ 的合适估计(要说明理由)

20. 设 X 取值 1 或 0，概率分布如下：

$$P_\theta(X=x)=\begin{cases}\theta^x(1-\theta)^{1-x}, & \text{当 }\theta\text{ 是有理数}\\(1-\theta)^x\theta^{1-x}, & \text{当 }\theta\text{ 是无理数}\end{cases}$$

其中 $\theta\in(0,1)$，$x=0,1$. 设 $X_1,X_2,\cdots,X_n$ 是 X 的样本，试找出 θ 的最大似然估计 $\hat{\theta}_n$，并证明 $\hat{\theta}_n$ 不具有相合性.

21. 设 $X_1,X_2,\cdots,X_n$ 是来自密度为

$$f(x;\theta)=\frac{1}{2}e^{-|x-\theta|}$$

的总体的样本，$-\infty<\theta<+\infty$，试求出 θ 的最大似然估计 $\hat{\theta}_n$ 并证明 $\hat{\theta}_n$ 是强相合的.

22. 设 $X_1,X_2,\cdots,X_n$ 是参数为 λ($\lambda>0$ 未知)的 Poisson 分布的样本，证明 $\overline{X}$ 是 λ 的强相合估计.

23. 设 $X_1,X_2,\cdots,X_n$ 是 $(0,\theta)$ 上的均匀分布的样本，求 θ 的置信水平为 $1-\alpha$ 的置信区间. 设得到了 5 个样本值 0.08，0.28，0.53，0.91，0.89，求 θ 的置信水平为 0.95 的置信区间.

24. 设 $X_1,X_2,\cdots,X_n$ 是 Bernoulli 分布 $b(1,p)$ 的样本，试根据中心极限定理构造 n 值较大时 p 的近似 $1-\alpha$ 置信水平的置信区间.

25. 设 $X_1,X_2,\cdots,X_n$ 是 $N(\mu,\sigma^2)$ 的样本，证明样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

的方差为

$$\mathrm{Var}(S^2)=\frac{2\sigma^4}{n-1}\quad(n\geqslant 2)$$

26. 设 $X_1,\cdots,X_n$ 独立同分布，共同分布函数是 $F(x)$，$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 是 $X_1,\cdots,X_n$ 的次序统计量. 证明：

(1) 对一切 $1\leqslant i\leqslant n$ 有

$$P(X_{(i)}\leqslant x)=\frac{n!}{(i-1)!(n-i)!}\int_0^{F(x)}u^{i-1}(1-u)^{n-i}\mathrm{d}u$$

(2) 若 $F(x)$ 是连续函数，则

$$E[F(X_{(i)})] = \frac{i}{n+1}$$

$$\text{Var}[F(X_{(i)})] = \frac{i(n+1-i)}{(n+1)^2(n+2)}$$

27. 用放射性同位素法可以客观地测定地层的年代. 为了研究这种方法的准确度,对同一地层取了 19 个样,测定了年代,数据如下(单位:百万年):

249 254 243 268 253 269 287 241 273 306 303

280 260 256 278 344 304 283 310

假设这些数据是正态分布的样本,求方差 σ^2 的置信水平为 0.95 的置信区间和标准差 σ 的置信水平为 0.95 的置信区间.

28. 科学中的伟大发现往往是由比较年轻的人提出的. 下表是 16 世纪中叶到 20 世纪的 12 个重大科学突破的情况:

科学发现	科学家	年份	年龄
日心说	哥白尼	1543	40
望远镜及天文学的基本定律	伽利略	1600	43
动力学、万有引力、微积分	牛顿	1665	23
电的实质	富兰克林	1746	40
燃烧即氧化	拉瓦锡	1774	31
地球的演变	莱尔	1830	33
进化论	达尔文	1858	49
光的电磁场特性	麦克斯韦	1864	33
放射性	居里	1896	34
量子力学	普朗克	1901	43
狭义相对论	爱因斯坦	1905	26
量子力学的数学基础	薛定谔	1926	39

设提出重大科学发现时科学家的年龄服从 $N(\mu,\sigma^2)$ 分布(μ,σ^2 未知),构造 μ 的置信水平为 0.95 的置信区间.

第三章 假设检验

§1 问题的提法

上一章我们介绍了估计的方法，但在实践中还有许多重要问题与估计问题的提法不同，也需要我们去解决. 请看下列简单的例子.

例 1.1 某厂有一批产品，共 200 件，须经检验合格才能出厂，按国家标准，次品率不得超过 3%. 今在其中任意抽取了 10 件，发现这 10 件中有 2 件是次品. 问这批产品能否出厂？

从直观上看，这批产品是不能出厂的，但理由何在？

设这批产品的次品率是 p，问题转化为：如何根据抽样的结果来判断“$p \leqslant 3\%$”成立与否？

例 1.2 用某仪器间接测量温度，重复测量 5 次，所得数据(℃)是 1 250，1 265，1 260，1 275，1 245，而用别的办法测得精确温度是 1 277(可看作温度的真值)，问此仪器间接测量有无系统偏差？

用 X 代表用这个仪器测得的温度值，当然这是一个随机变量，得到的 5 个数据是 X 的一个样本. 问题化为：如何判断等式“$EX = 1\,277$”成立与否？

例 1.3 在针织品的漂白工艺过程中，要考察温度对针织品断裂强力(主要质量指标)的影响. 为了比较 70 ℃与 80 ℃的影响有无差别. 在这两个温度下，分别重复作了 8 次试验，得数据如下：

70 ℃时的强力：20.5，18.8，19.8，20.9，21.5，19.5，21.0，21.2

80 ℃时的强力：17.7，20.3，20.0，18.8，19.0，20.1，20.2，19.1

究竟 70 ℃时的强力与 80 ℃时的强力有没有差别? 用 X 表示 70 ℃时的强力,Y 表示 80 ℃时的强力,问题变成:如何判断 X 与 Y 是否遵从同样的分布?

例 1.4 为研究抽烟与慢性支气管炎是否有关,调查了 n 个人,数据按照是否抽烟与是否患病(慢性支气管炎)可列为下面 2×2表

	有病	无病
抽烟	n_{11}	n_{12}
不抽烟	n_{21}	n_{22}

令
$$X=\begin{cases}1, & 若抽烟\\ 0, & 若不抽烟\end{cases},\quad Y=\begin{cases}1, & 若有病\\ 0, & 若无病\end{cases}$$
问题化为判断"X 与 Y 相互独立"是否成立.

例 1.5 怎样根据一个随机变量的样本值,判断该随机变量是否服从正态分布 $N(\mu,\sigma^2)$?

更一般地,如何根据样本去判断随机变量以给定的函数 $F_0(x)$为其分布函数?

这些例子所代表的问题是很广泛的. 其共同点是从样本值出发去判断关于总体分布的一个"看法"是否成立. 例 1.1 的看法是"次品率 $p\leqslant 0.03$",例 1.2 是"$EX=1\,277$",例 1.3 的看法是"X 与 Y 有相同的分布",例 1.4 的看法是"X 与 Y 相互独立",例 1.5 则是"X 的分布函数是 $F_0(x)$".

"看法"又叫做"假设".

这些就是所谓假设检验问题(或叫假设的鉴定问题).

本章介绍假设检验的基本理论知识和一些常用的检验方法,判断所关心的"假设"是否成立.

例 1.1、例 1.2 和例 1.5 中"假设"都是关于一个随机变量的分布的判断,这叫做一个总体的检验问题,例 1.3 和例 1.4 的"假设"是关于两个随机变量的分布的判断,这叫做二总体的检验问

题.也可以考虑三个或更多个总体的检验问题.

我们主要论述单总体的检验问题.把我们要检验的"假设"记作 H_0(通常叫做零假设),H_0 是关于随机变量(总体)的分布的一个"看法".说得更确切些,设 X 的分布函数为 $F(x,\theta)$,其中 θ 属于 Θ,这里 Θ 是一个已知集合(是实数集合,或向量组成之集合,也可以是更一般的符号组成之集合).零假设 H_0 通常可以表示成这样的形式:$\theta\in\Theta_0$,这里 Θ_0 是 Θ 之非空真子集.通常把"$\theta\in\Theta-\Theta_0$"叫做对立假设(或叫备择假设),记作 H_a 或 H_1.

怎样根据样本值对 H_0 进行检验呢?这就需要对"检验法"给出合理的定义.直观上说,所谓一个检验法,就是给出一个规则,对给定的样本值 $x_1,\cdots,x_n$ 进行明确表态:接受假设 H_0 还是拒绝假设 $H_0$①.

用数学语言可以说得更清楚些,设 S 是所有可能的样本值 $(x_1,\cdots,x_n)$(n 固定)组成的集合(样本空间),所谓一个检验法就是指空间 S 的一个划分:$S=S_1\cup S_2$(S_1 与 S_2 不相交).当 $(x_1,\cdots,x_n)\in S_1$ 时接受假设 H_0;当 $(x_1,\cdots,x_n)\in S_2$ 时拒绝 H_0,这 S_1 叫接收域,S_2 叫否定域.因为 $S_1=S-S_2$,故只要知道了否定域,就知道了检验法.每个检验法对应一个否定域;反之,任给定 S 的一个子集 W,则有一个检验法以 W 作为它的否定域.故研究检验法就相当于研究否定域.为便于数学处理,通常要求否定域是样本空间 S 中的 Borel 集.S 中的 Borel 集太多了,究竟应选哪一个对检验 H_0 是最合适的呢?这就涉及检验法优良性的标准.为了分析这一问题,我们看看在取定一个否定域 W(即选定一个检验法)后,有什么后果.

零假设 H_0 在客观上只有两种可能性:真、假.样本值 $(x_1,\cdots,x_n)$ 也只有两种可能性:属于否定域 W、不属于 W.若采

---

① 本章只研究这种固定样本量的检验法.至于样本量不固定的情况,现代也有大量研究,归属于数理统计学的一个分支——序贯分析,见第六章.

用 W 作否定域,则在观察到样本值$(x_1,\cdots,x_n)$时只可能有下列四种情况:

(1) H_0 真,而$(x_1,\cdots,x_n)$属于 W;

(2) H_0 真,而$(x_1,\cdots,x_n)$不属于 W;

(3) H_0 假,而$(x_1,\cdots,x_n)$属于 W;

(4) H_0 假,而$(x_1,\cdots,x_n)$不属于 W.

根据我们的规则,在情形(1)应拒绝 H_0,在情形(2)应接受 H_0,在情形(3)应拒绝 H_0,在情形(4)应接受 H_0. 情形(2)、(3)当然好,对 H_0 的表态与客观实际相符. 但(1)、(4)两种情形下,表态犯了错误:与客观实际不符. 在情形(1)下出现的错误是把本来真实的看法 H_0 进行了否定,这种"以真为假"的错误叫做第一类错误. 在情形(4)下出现的错误是把本来虚假的看法 H_0 接受下来,这种"以假为真"的错误叫做第二类错误. 由于样本值有随机性,这两类错误一般难以避免.

样本值 $x_1,\cdots,x_n$ 乃是样本 $X_1,\cdots,X_n$ 的值,在不引起误会的情况下,同一个字母有时表示随机变量,有时表示随机变量的值,这从上下文看不会引起误会. 为了研究两类错误,引进一些概念和符号.

用 $P(A|\theta)$表示 X 的分布的"参数"的真值是 θ 时事件 A 发生的概率(或用记号 $P_\theta(A)$).

$$
\begin{aligned}
L_W(\theta) &\overset{\mathrm{d}}{=\!=} P(\text{接受 } H_0 \mid \theta) \\
&= P((X_1,\cdots,X_n) \overline{\in} W \mid \theta)
\end{aligned}
$$

$$
\begin{aligned}
\rho_W(\theta) &\overset{\mathrm{d}}{=\!=} P(\text{拒绝 } H_0 \mid \theta) \\
&= P((X_1,\cdots,X_n) \in W \mid \theta)
\end{aligned}
$$

$L_W(\theta)$叫做检验法(否定域)的操作特性函数(简称 OC 函数),$\rho_W(\theta)$叫做 W 的功效函数. 显然 $L_W(\theta)=1-\rho_W(\theta)$.

当 $\theta\in\Theta_0$ 时,$\rho_W(\theta)$表示犯第一类错误的概率. 当 $\theta\overline{\in}\Theta_0$ 时,$1-\rho_W(\theta)$表示犯第二类错误的概率. 我们自然希望选取这样的否

定域，它犯两类错误的概率都很小．换言之，我们要求功效函数 $\rho_W(\theta)$ 当 $\theta\in\Theta_0$ 时很小，当 $\theta\overline{\in}\Theta_0$ 时很大．

遗憾的是，对给定的样本量 n 来讲，一般而论，第一类错误的概率小时，第二类错误的概率就大，反之亦然．因而不能做到两类错误的概率都任意小．所以问题的正确提法是：应选 W 使两类错误的概率都尽可能的小．

通常称 $\sup\limits_{\theta\in\Theta_0}\rho_W(\theta)$ 为 W 的检验水平①（也称显著性水平简称水平）．在实际工作中常常这样提出问题：对给定的小正数 α（通常取 $\alpha=0.10, 0.05$ 或 0.01），如何从检验水平不超过 α 的所有否定域中选取犯第二类错误的概率尽可能小的否定域？

我们常把要考虑的检验问题用下列符号表示

$$H_0:\theta\in\Theta_0\leftrightarrow H_\alpha:\theta\in\Theta_1$$

这里 H_0 是待检验的"假设"，H_α 是备择假设，Θ_1 与 Θ_0 不相交．在很多情形下，$\Theta_1=\Theta-\Theta_0$．

定义 1.1 称 W 是检验水平为 α 的一致最大功效的否定域（简称 UMP 否定域），若 W 的水平为 α 而且对一切水平不超过 α 的否定域 $\widetilde{W}$ 均有

$$\rho_W(\theta)\geqslant\rho_{\widetilde{W}}(\theta)\quad(\text{一切 }\theta\in\Theta_1)$$

能找出一致最大功效的否定域当然好，可惜这种否定域难得存在．比"一致最大功效"较弱的要求是"一致最大功效无偏"．

定义 1.2 称 W 是检验水平为 α 的无偏否定域，若对一切 $\theta\in\Theta_1$，有

$$\rho_W(\theta)\geqslant\alpha$$

无偏性是很自然的要求："假设"在它真实时遭拒绝的概率不大于它虚假时遭拒绝的概率．

---

① 有时称 $\sup\limits_{\theta\in\Theta_0}\rho_W(\theta)$ 为 W 的精确检验水平．给定正数 $\alpha(0<\alpha<1)$，精确检验水平恰为 α 的否定域 W 不一定存在．当 W 的精确检验水平不超过 α 时，有时也说 W 的检验水平是 α．

定义 1.3 称 W 是水平为 α 的一致最大功效无偏否定域，若 W 是水平为 α 的无偏否定域，而且对任何水平为 α 的无偏否定域 $\widetilde{W}$ 恒有

$$\rho_W(\theta) \geqslant \rho_{\widetilde{W}}(\theta) \quad (\text{一切 } \theta \in \Theta_1)$$

以后将会看到，大量常用的否定域是一致最大功效无偏的(UMPU).

在以上定义里“否定域”一词均可用“检验法”或“检验”代替.

寻找一致最大功效(或一致最大功效无偏)的否定域(检验法)是相当复杂的问题，需要对零假设 H_0 的具体结构及总体 X 的概率性质进行具体分析. 在许多情形下已有很好的答案，还有许多情形尚未研究清楚或答案不够令人满意. 从§2 起我们对一些重要的特殊情形进行讨论. 为了叙述明确起见，在一些定理的叙述中均假定随机变量是连续型的，读者不难验证，这些定理的结论可推广至离散型随机变量，证明方法基本不变，只需将积分号换成求和号.

作为本节的结尾，我们指出假设检验的接受域与置信区间有一种简单而深刻的联系. 设 X 的分布函数是 $F(x,\theta)$，θ 是未知参数，$\underset{\sim}{X}=(X_1,\cdots,X_n)$是 X 的样本. 考虑检验问题：

$$H_0: \theta=\theta_0 \leftrightarrow H_1: \theta \neq \theta_0$$

设 $A(\theta_0)$是 H_0 的接受域(检验水平不超过 α)，即当且仅当样本 $\underset{\sim}{X}$ 的值属于$A(\theta_0)$时接受假设 H_0，且

$$P(\underset{\sim}{X} \overline{\in} A(\theta_0) \mid \theta_0) \leqslant \alpha$$

令 $S(\underset{\sim}{X})=\{\theta: \underset{\sim}{X} \in A(\theta)\}$，则对一切 θ 有

$$P(\theta \in S(\underset{\sim}{X}) \mid \theta) \geqslant 1-\alpha$$

由此可见，如果 $S(\underset{\sim}{X})$是区间，则它就是 θ 的置信水平为 $1-\alpha$ 的置信区间.

例 1.6 设 $X \sim N(\theta,1)$，$\theta \in (-\infty,\infty)$. $\underset{\sim}{X}=(X_1,\cdots,X_n)$是

X 的样本. 对于检验问题:

$$\theta=\theta_0 \leftrightarrow \theta \neq \theta_0$$

可采用接受域:

$$A(\theta_0)=\{(x_1,\cdots,x_n):|\bar{x}-\theta_0|\leqslant c\}$$

其中 $\bar{x}=\frac{1}{n}\sum_{1}^{n} x_i$, c 满足

$$P(\underset{\sim}{X}\overline{\in}A(\theta_0)|\theta_0)=\alpha$$

(α 是检验水平). 易知 $c=\frac{1}{\sqrt{n}}z_{1-\frac{\alpha}{2}}$ $\left(z_{1-\frac{\alpha}{2}}\text{ 是标准正态分布的 } 1-\frac{\alpha}{2}\right.$ 分位数$\Big)$. 不难看出, $S(\underset{\sim}{X})=\{\theta:\underset{\sim}{X}\in A(\theta)\}=\{\theta:|\overline{X}-\theta|\leqslant c\}=$ $[\overline{X}-c,\overline{X}+c]$, 这里 $\overline{X}=\frac{1}{n}\sum_{1}^{n}X_i$. 这表明 $[\overline{X}-c,\overline{X}+c]$ 是 θ 的 $1-\alpha$ 水平置信区间.

以后将看到, 在许多具体的假设检验问题中常可选取具有某种优良性的接受域, 因而根据接受域构造的置信区间也有其优良之处.

§2 N-P 引理及似然比检验法

我们首先考虑最简单的情形. 设 $X\sim F(x,\theta)$, $\theta\in\Theta=\{\theta_1,\theta_2\}$ $(\theta_1\neq\theta_2)$, 设 X 是连续型随机变量, 密度函数是 $f(x,\theta)$ (对离散型随机变量可进行类似的讨论). $\Theta_0=\{\theta_1\}$. 检验问题是:

$$H_0:\theta=\theta_1 \leftrightarrow H_a:\theta=\theta_2$$

设 X 的样本是 $X_1,\cdots,X_n$, 应怎样确定最好的否定域呢? 记

$$\underset{\sim}{x}=(x_1,\cdots,x_n),\quad d\underset{\sim}{x}=dx_1dx_2\cdots dx_n,\quad L(\underset{\sim}{x},\theta)=\prod_{i=1}^{n}f(x_i,\theta).$$

定理 2.1(Neyman - Pearson 引理) 给定 $\alpha\in(0,1)$, 设

$$W_0=\{\underset{\sim}{x}:L(\underset{\sim}{x},\theta_2)>\lambda_0 L(\underset{\sim}{x},\theta_1)\} \tag{2.1}$$

(这里 $\lambda_0\geqslant 0$)适合

$$\int_{W_0}\cdots\int L(\underset{\sim}{x},\theta_1)\mathrm{d}\,\underset{\sim}{x}=\alpha \tag{2.2}$$

则对任何否定域 $W\subset R^n$，只要 $\rho_W(\theta_1)\leqslant\alpha$，就一定有

$$\rho_{W_0}(\theta_2)\geqslant\rho_W(\theta_2)$$

换句话说，W_0 是所有检验水平不超过 α 的否定域中犯第二类错误的概率最小的一个.

证明：设 W 是任何满足 $\rho_W(\theta_1)\leqslant\alpha$ 的否定域，则

$$\begin{aligned}
&\rho_{W_0}(\theta_2)-\rho_W(\theta_2)\\
&=P((X_1,\cdots,X_n)\in W_0|\theta_2)-P((X_1,\cdots,X_n)\in W\mid\theta_2)\\
&=\int_{W_0}\cdots\int L(\underset{\sim}{x},\theta_2)\mathrm{d}\,\underset{\sim}{x}-\int_W\cdots\int L(\underset{\sim}{x},\theta_2)\mathrm{d}\,\underset{\sim}{x}\\
&=\int_{W_0-W}\cdots\int L(\underset{\sim}{x},\theta_2)\mathrm{d}\,\underset{\sim}{x}-\int_{W-W_0}\cdots\int L(\underset{\sim}{x},\theta_2)\mathrm{d}\,\underset{\sim}{x}\\
&\geqslant\lambda_0\left[\int_{W_0-W}\cdots\int L(\underset{\sim}{x},\theta_1)\mathrm{d}\,\underset{\sim}{x}-\int_{W-W_0}\cdots\int L(\underset{\sim}{x},\theta_1)\mathrm{d}\,\underset{\sim}{x}\right]\\
&=\lambda_0\left[\int_{W_0}\cdots\int L(\underset{\sim}{x},\theta_1)\mathrm{d}\,\underset{\sim}{x}-\int_W\cdots\int L(\underset{\sim}{x},\theta_1)\mathrm{d}\,\underset{\sim}{x}\right]\\
&=\lambda_0[\alpha-\rho_W(\theta_1)]\geqslant 0
\end{aligned}$$

证毕.

定理 2.1 乃是统计学中著名的 N－P 引理，它断言否定域(2.1)具有最优性(犯第二类错误的概率最小). 这个否定域可表示为：

$$W_0=\{\underset{\sim}{x}:\lambda>\lambda_0\} \tag{2.3}$$

其中 $\lambda=\lambda(\underset{\sim}{x})=L(\underset{\sim}{x},\theta_2)/L(\underset{\sim}{x},\theta_1)$ 叫做似然比. 这个否定域确定的检验法叫做似然比检验法. 满足(2.2)的 λ_0 是否存在呢? 当似

然比 $\lambda(\underset{\sim}{X})$ 的分布函数连续时，不难看出确有 λ_0 满足(2.2). ①

N-P 引理告诉我们，似然比检验法具有最优性. 自然问：是否还有别的检验法与似然比检验法一样优？分析定理 2.1 的证明过程不难得到下列唯一性结论.

定理 2.2 设 X 的分布密度是 $f(x,\theta_i)(i=1,2)$，X 的可能值集合 $\mathscr{X}=\{x: f(x,\theta_i)>0\}$ 与 i 无关. $\lambda(\underset{\sim}{x})=L(\underset{\sim}{x},\theta_2)/L(\underset{\sim}{x},\theta_1)$ $(\underset{\sim}{x}\in\mathscr{X}^n)$，设 $\underset{\sim}{X}=(X_1,\cdots,X_n)$ 是 X 的样本，若 $\lambda(\underset{\sim}{X})$ 在 θ_1 下的分布函数是连续的，则对任何 $\alpha\in(0,1)$，存在 $\lambda_0>0$，使得 $W_0=\{\underset{\sim}{x}:\lambda(\underset{\sim}{x})>\lambda_0\}$ 是水平为 α 的唯一最大功效的否定域. 这里"唯一"

---

① 当 $\lambda(\underset{\sim}{X})$ 的分布函数不连续，尤其 X 本身就是离散型随机变量时，可能不存在这样的 λ_0.

例如，X 的分布密度为

$$f(x,\theta_1)=\begin{cases}\dfrac{1}{3}, & \text{当 } x\in[0,1],\\ \dfrac{2}{3}, & \text{当 } x\in(1,2],\\ 0, & \text{其他}.\end{cases}$$

$$f(x,\theta_2)=\begin{cases}\dfrac{2}{3}, & \text{当 } x\in[0,1],\\ \dfrac{1}{3}, & \text{当 } x\in(1,2],\\ 0, & \text{其他}.\end{cases}$$

若样本量 $n=1$，则似然比为

$$\lambda(X)=\begin{cases}2, & \text{当 } X\in[0,1],\\ \dfrac{1}{2}, & \text{当 } X\in(1,2],\\ \text{任意值}, & \text{其他}.\end{cases}$$

于是对任何 $\lambda_0\geqslant 0$，

$$\int_{\{x:\lambda(x)>\lambda_0\}} f(x,\theta_1)\mathrm{d}x=\begin{cases}0, & \text{当 } \lambda_0\geqslant 2\\ \dfrac{1}{3}, & \text{当 } \lambda_0\in\left[\dfrac{1}{2},2\right)\\ 1, & \text{当 } \lambda_0<\dfrac{1}{2}\end{cases}$$

可见 $\alpha\neq\dfrac{1}{3}$ 时没有 λ_0 满足(2.2).

的含义是：若W也是水平为α的最大功效的否定域，则$\mu(W\triangle W_0)=0$（μ是 Lebesgue 测度，$W\triangle W_0\overset{\mathrm{d}}{=}(W-W_0)\cup(W_0-W)$）.

证明：由于$\lambda(\underset{\sim}{X})$的分布函数连续，故有$\lambda_0>0$，使得

$$P(\lambda(\underset{\sim}{X})>\lambda_0\,|\,\theta_1)=\alpha$$

于是W_0的水平是α. 设W是$\mathscr{X}^n$的任一子集，检验水平不超过α，即$P(\underset{\sim}{X}\in W\,|\,\theta_1)\leqslant\alpha$. 我们来证明：若$\mu(W\triangle W_0)>0$则必有

$$P(\underset{\sim}{X}\in W_0\,|\,\theta_2)>P(\underset{\sim}{X}\in W\,|\,\theta_2) \tag{2.4}$$

实际上

$$\begin{aligned}&P(\underset{\sim}{X}\in W_0\,|\,\theta_2)-P(\underset{\sim}{X}\in W\,|\,\theta_2)\\&=\int_{W_0}L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}-\int_{W}L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}\\&=\int_{W_0-W}L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}-\int_{W-W_0}L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}\end{aligned}$$

这里及下面我们恒用$\int$代替$\int\cdots\int$. 分两种情况讨论.

（一）$\boldsymbol{\mu(W_0-W)>0}$

此时在集合W_0-W上$L(\underset{\sim}{x},\theta_2)>\lambda_0L(\underset{\sim}{x},\theta_1)$于是

$$\int_{W_0-W}L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}>\lambda_0\int_{W_0-W}L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}$$

但在$W-W_0$上$L(\underset{\sim}{x},\theta_2)\leqslant\lambda_0L(\underset{\sim}{x},\theta_1)$，故

$$\begin{aligned}&P(\underset{\sim}{X}\in W_0\,|\,\theta_2)-P(\underset{\sim}{X}\in W\,|\,\theta_2)\\&>\lambda_0\left[\int_{W_0-W}L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}-\int_{W-W_0}L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}\right]\\&=\lambda_0[P(\underset{\sim}{X}\in W_0\,|\,\theta_1)-P(\underset{\sim}{X}\in W\,|\,\theta_1)]\geqslant0\end{aligned}$$

故(2.4)式成立.

(二) $\mu(W-W_0)>0$

令

$$D=\{\underset{\sim}{x}:\underset{\sim}{x}\in\mathscr{X}^n, L(\underset{\sim}{x},\theta_2)=\lambda_0 L(\underset{\sim}{x},\theta_1)\}$$

因为 $\lambda(\underset{\sim}{X})$ 的分布函数连续，故 $P(\underset{\sim}{X}\in D|\theta_1)=0$，即

$$\int_D L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}=0$$

但 $L(\underset{\sim}{x},\theta_1)>0$，故 $\mu(D)=0$. 从而

$$\int_{W-W_0} L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}=\int_{W-(W_0\cup D)} L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}$$

$$<\int_{W-W_0}\lambda_0 L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}$$

由此不难推知(2.4)仍然成立. 证毕.

我们还指出，上述似然比检验法还是无偏的.

定理 2.3 在定理 2.1 的假定下

$$\rho_{W_0}(\theta_2)\geqslant\alpha=\rho_{W_0}(\theta_1)$$

证明：记 $W_0^C=\mathscr{X}^n-W_0$，则

$$\begin{aligned}
\rho_{W_0}(\theta_2)-\alpha &=\int_{W_0} L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}-\alpha \\
&=\int_{W_0}(1-\alpha)L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x}-\alpha\int_{W_0^C} L(\underset{\sim}{x},\theta_2)\mathrm{d}\underset{\sim}{x} \\
&\geqslant\lambda_0(1-\alpha)\int_{W_0} L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}-\alpha\lambda_0\int_{W_0^C} L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x} \\
&=\lambda_0\left[\int_{W_0} L(\underset{\sim}{x},\theta_1)\mathrm{d}\underset{\sim}{x}-\alpha\right] \\
&=\lambda_0(\alpha-\alpha)=0
\end{aligned}$$

证毕.

现举例说明如何应用 N－P 引理.

例 2.1 设 $X\sim N(\mu,1)$，$\mu\in\Theta=\{0,2\}$，检验问题是：

$$H_0:\mu=0\leftrightarrow H_a:\mu=2$$

给定 $\alpha=0.05$，设样本是 $X_1,\cdots X_n$，我们来求出最大功效的否

定域.

X 的密度函数是

$$f(x,\mu) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\mu)^2\right\}$$

根据 N－P 引理只须找$\{(x_1,\cdots,x_n):\lambda>\lambda_0\}$型的否定域，其中

$$\lambda = \prod_1^n f(x_i,2)\Big/\prod_1^n f(x_i,0)$$

易知 $\lambda=e^{2n\overline{X}-2n}$，为了 $\lambda>\lambda_0$ 必须且只须 $\overline{X}>\frac{\ln\lambda_0}{2n}+1$，记 $C=\frac{\ln\lambda_0}{2n}+1$，故应选 C 满足 $P(\overline{X}>C|0)=0.05$，这里 $\overline{X}=\frac{1}{n}\sum_1^n X_i\sim N\left(0,\frac{1}{n}\right)$（当 $\mu=0$ 时），于是

$$\sqrt{n}\,\overline{X}\sim N(0,1)$$

所以 C 应满足：

$$\int_{\sqrt{n}C}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=0.05$$

查表知 $\sqrt{n}C\doteq 1.65$，$C=\frac{1.65}{\sqrt{n}}$，最大功效的否定域是 $W_0=\left\{(x_1,\cdots,x_n):\overline{x}>\frac{1.65}{\sqrt{n}}\right\}$，其中 $\overline{x}=\frac{1}{n}\sum_{i=1}^n x_i$. 这个否定域是由统计量 $\overline{X}$ 来确定的. 以后将看到，在许多其他的假设检验问题中也有类似情况，优良的否定域往往可以利用直观意义很清楚的统计量来确定.

§3 单参数情形的假设检验

设 X 的分布函数是 $F(x,\theta)$，$\theta\in\Theta=(a,b)$，$-\infty\leqslant a<b\leqslant\infty$ 这就是单个实参数的情形. 在实际工作中遇到的多是下列类型的检验问题：

(1) $H_0:\theta\leqslant\theta_1 \leftrightarrow H_a:\theta>\theta_1$

(或 $H_0:\theta\geqslant\theta_1 \leftrightarrow H_a:\theta<\theta_1$)

(2) $H_0:\theta\overline{\in}(\theta_1,\theta_2) \leftrightarrow H_a:\theta\in(\theta_1,\theta_2)$

(3) $H_0:\theta\in[\theta_1,\theta_2] \leftrightarrow H_a:\theta\overline{\in}[\theta_1,\theta_2]$

(4) $H_0:\theta=\theta_1 \leftrightarrow H_a:\theta\neq\theta_1$

其中 $\theta_1<\theta_2,\theta_1\in(a,b),\theta_2\in(a,b)$.

本节只讨论单参数指数型分布的检验问题.

定义 3.1 设 X 的可能值集合是 $\mathscr{X}$. 称 X 服从单参数指数型分布,若 X 的分布密度(或概率函数)$f(x,\theta)$有下列表达式:

$$f(x,\theta)=S(\theta)h(x)\mathrm{e}^{Q(\theta)V(x)} \tag{3.1}$$

其中 $\theta\in\Theta=(a,b),S(\theta)>0,x\in\mathscr{X},h(x)>0,Q(\theta)$是 θ 的严格增函数(参看第二章例 2.3).

指数型分布十分广泛,指数分布、二项分布、负二项分布、Poisson 分布、均值已知或方差已知的正态分布等等都是单参数指数型的.

为了研究复杂假设的检验,我们再对上节例 2.1 考察一下. 有人可能注意到,该例里 $\Theta=\{0,2\}$,但导出的具有最大功效的否定域与“2”无关,即若将 Θ 变与 $\Theta=\{0,\mu_1\}(\mu_1>0)$,则导出的具有最大功效的否定域与 μ_1 无关. 其原因是第一类错误与 μ_1 无关,而 μ_1 的作用体现在第二类错误上,犯第二类错误的概率是 $1-\rho_{W_0}(\mu_1)=1-P_{\mu_1}\left(\overline{X}>\frac{1.65}{\sqrt{n}}\right)=\Phi(1.65-\sqrt{n}\mu_1)$($\Phi$ 是标准正态分布函数). 可以看出,给定 n 时 μ_1 越大则犯第二类错误的概率越小;μ_1 越小则犯第二类错误的概率越大,即越不易分辨 0 与 μ_1.

上述事实提示我们,既然 W_0 与 μ_1 无关,例 2.1 中得到的 W_0 应该是针对零假设 $H_0:\mu=0$ 与备择假设 $H_a:\mu>0$ 的 UMP 检验. 更进一步,还可以猜想,当检验问题变为:$H_0:\mu\leqslant0\leftrightarrow H_a:\mu>0$时,$W_0$ 仍是 UMP 检验。下面的定理 3.1 表明,上述猜想不仅是正确的,而且可以由正态分布扩展到单参数指数型分布的情形.

为确定计，以下只讨论指数型分布有密度 $f(x,\theta)$ 的情形，对于离散型情形可进行类似的讨论.

设 $X_1,\cdots,X_n$ 是样本，$t(X_1,\cdots,X_n)=\sum_{i=1}^{n}V(X_i)$，$\underset{\sim}{X}=(X_1,\cdots,X_n)$，$L(\underset{\sim}{X},\theta)=\prod_{j=1}^{n}f(X_j,\theta)$.

定理 3.1 设 X 的分布密度有(3.1)型. 给定检验问题

$$H_0:\theta\leqslant\theta_1\leftrightarrow H_a:\theta>\theta_1$$

对 $\alpha\in(0,1)$，若存在 C 满足

$$P\left(\sum_{1}^{n}V(X_i)>C|\theta_1\right)=\alpha \tag{3.2}$$

则

$$W_0=\left\{(x_1,\cdots,x_n):\sum_{1}^{n}V(x_i)>C\right\} \tag{3.3}$$

是检验水平为 α 的一致最大功效的否定域.

证明：从(3.2)知 $P(\underset{\sim}{X}\in W_0|\theta_1)=\alpha$. 下面证明：对一切 $\theta_0<\theta_1$，有 $P(\underset{\sim}{X}\in W_0|\theta_0)\leqslant\alpha$，为此研究检验问题：

$$\hat{H}_0:\theta=\theta_0\leftrightarrow\hat{H}_a:\theta=\theta_1 \tag{3.4}$$

记

$$\alpha_1\overset{\mathrm{d}}{=\!=}P(\underset{\sim}{X}\in W_0|\theta_0)$$

易知

$$\lambda_{\theta_0\theta_1}\overset{\mathrm{d}}{=\!=}\frac{L(\underset{\sim}{x},\theta_1)}{L(\underset{\sim}{x},\theta_0)}=\frac{S^n(\theta_1)}{S^n(\theta_0)}\mathrm{e}^{[Q(\theta_1)-Q(\theta_0)]\sum_{1}^{n}V(x_i)}$$

故

$$W_0\overset{\mathrm{d}}{=\!=}\left\{\underset{\sim}{x}:\sum_{1}^{n}V(x_i)>C\right\}=\{\lambda_{\theta_0\theta_1}>\lambda_0\}\quad(\underset{\sim}{x}=(x_1,\cdots,x_n))$$

根据 N－P 引理，W_0 是检验问题(3.4)的水平为 α_1 的最大功效的

否定域，根据定理 2.3，W_0 具有无偏性，即有

$$P(\underset{\sim}{X}\in W_0|\theta_1)\geqslant P(\underset{\sim}{X}\in W_0|\theta_0)=\alpha_1$$

这就证明了对一切 $\theta_0<\theta_1$，

$$P(\underset{\sim}{X}\in W_0|\theta_0)\leqslant\alpha$$

我们再来证：对任何水平不超过 α 的否定域 W，恒有

$$\rho_{W_0}(\theta)\geqslant\rho_W(\theta)\quad(\text{一切 }\theta>\theta_1)$$

为此，任意给定 $\theta_2>\theta_1$，研究检验问题：

$$\widetilde{H}_0:\theta=\theta_1\leftrightarrow\widetilde{H}_a:\theta=\theta_2\tag{3.5}$$

记

$$\lambda_{\theta_1\theta_2}\overset{\text{d}}{=}L(\underset{\sim}{x},\theta_2)/L(\underset{\sim}{x},\theta_1)$$

易知存在 λ_0 使得

$$\begin{aligned}W_0&\overset{\text{d}}{=}\left\{\underset{\sim}{x}:\sum_1^n V(x_i)>C\right\}\\&=\{\underset{\sim}{x}:\lambda_{\theta_1\theta_2}>\lambda_0\}\end{aligned}$$

根据 N－P 引理，W_0 是检验问题(3.5)的水平为 α 的最大功效否定域，既然 $\rho_W(\theta_1)\leqslant\alpha$，故

$$\rho_{W_0}(\theta_2)\geqslant\rho_W(\theta_2)$$

这就证明了 W_0 是水平为 α 的一致最大功效(UMP)否定域.

证毕.

对于检验问题：$H_0:\theta\geqslant\theta_1\leftrightarrow H_a:\theta<\theta_1$ 可进行类似的讨论.

现在应用定理 3.1 研究正态分布的两个检验问题.

(1) 设 $X\sim N(\mu,\sigma_0^2)$(σ_0 已知). 检验问题是

$$H_0:\mu\leqslant\mu_0\leftrightarrow H_a:\mu>\mu_0$$

其中 μ_0 是已知的. 设样本值是 $x_1,\cdots,x_n$，如何检验 H_0？

此时

$$\begin{aligned}f(x,\theta)&=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{1}{2\sigma_0^2}(x-\mu)^2\right\}\\&=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{\mu^2}{2\sigma_0^2}\right\}\cdot\exp\left\{-\frac{x^2}{2\sigma_0^2}\right\}\cdot\exp\left\{\frac{\mu x}{\sigma_0^2}\right\}\end{aligned}$$

对比(3.1)知 $V(x)=x$，依定理 3.1 知一致最大功效的否定域是

$$W_0=\left\{(x_1,\cdots,x_n):\sum_{i=1}^{n}x_i>C\right\} \tag{3.6}$$

其中 C 满足

$$P\left(\sum_{i=1}^{n}X_i>C|\mu_0\right)=\alpha$$

但 $\mu=\mu_0$ 时，

$$\overline{X}\sim N\left(\mu_0,\frac{\sigma_0^2}{n}\right)$$

故

$$P\left(\sum_{1}^{n}X_i>C|\mu_0\right)=P\left(\overline{X}>\frac{C}{n}|\mu_0\right)$$

$$=1-\Phi\left[\frac{Cn^{-\frac{1}{2}}-\sqrt{n}\mu_0}{\sigma_0}\right]$$

若 $\alpha=0.05$，查表知 $\Phi(1.65)\doteq0.95$. 于是

$$Cn^{-\frac{1}{2}}-\sqrt{n}\mu_0=1.65\sigma_0$$

故

$$C=n\mu_0+1.65\sigma_0\sqrt{n}$$

这时否定域(3.6)可以写成：

$$W_0=\left\{(x_1,\cdots,x_n):\bar{x}>\mu_0+\frac{1.65}{\sqrt{n}}\sigma_0\right\} \tag{3.7}$$

例 3.1 根据长期的经验和资料表明，某砖瓦厂所生产砖的“抗断强度”X 服从正态分布，方差 $\sigma^2=1.21$，今从该厂所生产的一批砖中，随机抽取 6 块，测得抗断强度数据如下：

32.56，29.66，31.64，30.00，31.87，31.03

问：这一批砖的平均抗断强度可否认为超过 30？

设 $X\sim N(\mu,\sigma^2)$；已知 $\sigma^2=1.21$，检验问题是

$$H_0:\mu\leqslant30\leftrightarrow H_a:\mu>30$$

设 $\alpha=0.05$，现在 $n=6$，$\mu_0=30$，$\sigma_0=1.1$，从(3.7)知否定域为

$$W_0=\{(x_1,\cdots,x_n):\bar{x}>30.74\}$$

现在根据所给的样本值 32.56,…,31.03 计算出 $\bar{x}=31.3>30.74$,故样本值落入了否定域,因而应否定 $H_0:\mu\leqslant 30$,所以可以认为这批砖的平均抗断强度超过 30.

(2) 设 $X\sim N(\mu_0,\sigma^2)$(μ_0 已知),设样本值是 $x_1,\cdots,x_n$,研究检验问题

$$H_0:\sigma^2\leqslant\sigma_0^2\leftrightarrow H_a:\sigma^2>\sigma_0^2$$

X 的密度函数是

$$f(x,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu_0)^2\right\}$$

对比(3.1)知 $V(x)=(x-\mu_0)^2$,从定理 3.1 知一致最大功效的否定域是

$$W_0=\left\{(x_1,\cdots,x_n):\sum_1^n(x_i-\mu_0)^2>C\right\}$$

其中 C 满足

$$P\left(\sum_{i=1}^n(X_i-\mu_0)^2>C|\sigma_0^2\right)=\alpha$$

但 $\sigma^2=\sigma_0^2$ 时,$\frac{1}{\sigma_0^2}\sum_{i=1}^n(X_i-\mu_0)^2$ 服从 n 个自由度的χ^2分布,利用χ^2分布的分位数表不难找出 C 来.

对于本节开始提出的假设检验问题 2)、3)、4),经过现代的深入研究,指数型分布情形已经完全解决了,见下列三个定理. 由于这些定理的数学证明较长,我们将在本节末尾的楷体字排印部分叙述证明,供有兴趣的读者参考.

定理 3.2 设 X 有分布密度

$$f(x,\theta)=S(\theta)h(x)e^{Q(\theta)V(x)}$$

这里 $S(\theta)>0,h(x)\geqslant 0,Q(\theta)$是 θ 的严格增函数.

给定检验问题:

$$H_0:\theta\overline{\in}(\theta_1,\theta_2)\leftrightarrow H_a:\theta\in(\theta_1,\theta_2)$$

设 $\underset{\sim}{X}=(X_1,\cdots,X_n)$是 X 的样本,且

$$W_0=\left\{(x_1,\cdots,x_n):C_1<\sum_{i=1}^{n}V(x_i)<C_2\right\}$$

若 $P(\underset{\sim}{X}\in W_0|\theta_i)=\alpha(i=1,2)$，则 W_0 是水平为 α 的一致最大功效的否定域(UMP 检验).

定理 3.3 设 X 有分布密度

$$f(x,\theta)=S(\theta)h(x)\mathrm{e}^{Q(\theta)V(x)}$$

$\theta\in(a,b)(-\infty\leqslant a<b\leqslant\infty)$，其中 $S(\theta)>0,h(x)\geqslant0,Q(\theta)$ 是 θ 的严格增连续函数.

给定检验问题：

$$H_0:\theta\in[\theta_1,\theta_2]\leftrightarrow H_a:\theta\overline{\in}[\theta_1,\theta_2]$$

$$(a<\theta_1<\theta_2<b)$$

设 $\underset{\sim}{X}=(X_1,\cdots,X_n)$ 是 X 的样本，且

$$W_0=\left\{(x_1,\cdots,x_n):\sum_{i=1}^{n}V(x_i)\text{ 小于 }C_1\text{ 或大于 }C_2\right\}$$

若 $C_1<C_2$ 使得

$$P(\underset{\sim}{X}\in W_0|\theta_i)=\alpha\quad(i=1,2)\tag{3.8}$$

则 W_0 是水平为 α 的一致最大功效的无偏否定域(UMPU 检验).

定理 3.4 设 X 有分布密度

$$f(x,\theta)=S(\theta)h(x)\mathrm{e}^{Q(\theta)V(x)}$$

这里 $\theta\in(a,b)(-\infty\leqslant a<b\leqslant\infty),S(\theta)>0,h(x)\geqslant0,Q'(\theta)>0.$

给定检验问题：

$$H_0:\theta=\theta_0\leftrightarrow H_a:\theta\neq\theta_0\quad(\theta_0\in(a,b))$$

设 $\underset{\sim}{X}=(\underset{\sim}{X}_1,\cdots,\underset{\sim}{X}_n)$ 是样本，$t(x_1,\cdots,x_n)=\sum_{i=1}^{n}V(x_i)$ 且

$$W_0=\{(x_1,\cdots,x_n):t(\underset{\sim}{x})<C_1\text{ 或}>C_2\}$$

若 $C_1<C_2$ 使得

$$P(\underset{\sim}{X}\in W_0|\theta_0)=\alpha\tag{3.9}$$

$$E_{\theta_0}(I_{W_0}(\underset{\sim}{X})t(\underset{\sim}{X}))=\alpha E_{\theta_0}t(\underset{\sim}{X})\tag{3.10}$$

（E_{θ_0} 乃是对应 $\theta=\theta_0$ 时的数学期望）.

则 W_0 是水平为 α 的一致最大功效无偏否定域（UMPU 检验）.

条件（3.10）什么时候得到满足呢？我们指出，若 $t(\underset{\sim}{X})=\sum_{i=1}^{n}V(X_i)$ 的分布关于某个数 r_0 是对称的，即

$$P(t<r_0-a\,|\,\theta_0)=P(t>r_0+a\,|\,\theta_0)\quad(\text{一切 } a\geqslant 0)$$

又 $W_0=\{\underset{\sim}{x}:t(\underset{\sim}{x})$ 小于 r_0-C 或者 $t(\underset{\sim}{x})$ 大于 $r_0+C\}$ 满足 $P(\underset{\sim}{X}\in W_0\,|\,\theta_0)=\alpha$，则（3.10）一定成立. 实际上，此时有

$$\begin{aligned}
&E_{\theta_0}(I_{W_0}(\underset{\sim}{X})t(\underset{\sim}{X}))\\
&=\int_{W_0}(t-r_0)\prod_{i=1}^{n}f(x_i,\theta_0)\mathrm{d}\underset{\sim}{x}+r_0\int_{W_0}\prod_{i=1}^{n}f(x_i,\theta_0)\mathrm{d}\underset{\sim}{x}\\
&=\int_{\{|t-r_0|>C\}}(t-r_0)\prod_{i=1}^{n}f(x_i,\theta_0)\mathrm{d}\underset{\sim}{x}+r_0P(\underset{\sim}{X}\in W_0\,|\,\theta_0)\\
&=0+r_0\alpha
\end{aligned}$$

由于 $E_{\theta_0}(t-r_0)=0$，故 $E_{\theta_0}(I_{W_0}(\underset{\sim}{X})t(\underset{\sim}{X}))=\alpha E_{\theta_0}t$. 这表明条件（3.10）成立.

例 3.2 设 $x_1,\cdots,x_n$ 是来自 $N(\theta,\sigma_0^2)$ 的样本，σ_0 已知，研究检验问题：

$$H_0:\theta=\theta_0\leftrightarrow H_a:\theta\neq\theta_0$$

注意密度

$$\begin{aligned}
f(x,\theta)&=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{1}{2\sigma_0^2}(x-\theta)^2\right\}\\
&=\frac{1}{\sqrt{2\pi\sigma_0^2}}\mathrm{e}^{-\frac{x^2}{2\sigma_0^2}-\frac{\theta^2}{2\sigma_0^2}}\cdot\mathrm{e}^{\frac{\theta}{\sigma_0^2}x}
\end{aligned}$$

对比（3.1）知 $V(x)=x$，故从定理 3.4 知一致最大功效无偏否定域是：

$$W_0=\left\{(x_1,\cdots,x_n):\sum_{i=1}^{n}x_i<C_1\text{ 或 }>C_2\right\}$$

这与直观是一致的，其中 C_1,C_2 满足：

$$P\left(\sum_{i=1}^{n} X_i < C_1 \text{ 或 } > C_2 \mid \theta_0\right) = \alpha$$

$$E_{\theta_0}\left(I_{W_0}(\underset{\sim}{X}) \cdot \sum_{i=1}^{n} X_i\right) = \alpha E_{\theta_0}\left(\sum_{i=1}^{n} X_i\right) \tag{3.11}$$

但 $\theta=\theta_0$ 时，$\sum_{i=1}^{n} X_i \sim N(n\theta_0, n\sigma_0^2)$，$\sum_{i=1}^{n} X_i$ 的分布关于 $r_0=n\theta_0$ 是对称的，故只要取 $C>0, C_1=n\theta_0-C, C_2=n\theta_0+C$ 使得

$$P\left(\sum_{i=1}^{n} X_i < C_1 \text{ 或 } > C_2 \mid \theta_0\right) = \alpha \tag{3.12}$$

则(3.11)一定成立.

从(3.12)知

$$\begin{aligned} \alpha &= P\left(\left|\sum_{i=1}^{n} X_i - n\theta_0\right| > C \mid \theta_0\right) \\ &= 2\Phi\left(\frac{-C}{\sqrt{n}\sigma_0}\right) \end{aligned}$$

于是

$$\frac{-C}{\sqrt{n}\sigma_0} = z_{\frac{\alpha}{2}}$$

这里 $z_{\frac{\alpha}{2}}$ 是标准正态分布的 $\frac{\alpha}{2}$ 分位数，满足

$$\int_{-\infty}^{z_{\frac{\alpha}{2}}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u = \frac{\alpha}{2}$$

所以

$$C = -\sqrt{n}\,\sigma_0 z_{\frac{\alpha}{2}}$$

从而

$$W_0 = \left\{(x_1, \cdots, x_n): \left|\sum_{1}^{n} x_i - n\theta_0\right| > \sqrt{n}\,\sigma_0 z_{1-\frac{\alpha}{2}}\right\}$$

(因为 $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$).

例 3.3 设 $X_1, \cdots, X_n$ 是来自 $N(\theta, \sigma_0^2)$ 的样本(σ_0 已知)，研究

检验问题：

$$H_0:\theta\in[\theta_1,\theta_2]\leftrightarrow H_a:\theta\overline{\in}[\theta_1,\theta_2]$$

从定理 3.3 知，一致最大功效无偏否定域是：

$$W_0=\left\{(x_1,\cdots,x_n):\sum_1^n x_i<C_1\text{ 或 }>C_2\right\}.$$

这里 C_1,C_2 满足

$$P((X_1,\cdots,X_n)\in W_0|\theta_i)=\alpha\quad(i=1,2)$$

即有

$$\Phi\left(\frac{C_2-n\theta_1}{\sqrt{n}\sigma_0}\right)-\Phi\left(\frac{C_1-n\theta_1}{\sqrt{n}\sigma_0}\right)=1-\alpha$$

$$\Phi\left(\frac{C_2-n\theta_2}{\sqrt{n}\sigma_0}\right)-\Phi\left(\frac{C_1-n\theta_2}{\sqrt{n}\sigma_0}\right)=1-\alpha$$

C_1,C_2 需通过正态分布表多次试查才能找出近似值.

当 $\theta_2=-\theta_1$ 时，必有 $C_2=-C_1$，此时 C_1 利用正态分布表较易求出.

为了给出定理 3.2，定理 3.3 和定理 3.4 的数学证明，我们首先要推广 N－P引理.

引理 3.1（N－P 引理的推广） 设 $f_0(x),f_1(x),f_2(x),\phi(x),\phi^*(x)$，$\bar{\phi}(x)$都是 S 上的可积函数（S 是 n 维空间中的区域或为更一般的集合），$0\leqslant\phi^*(x)\leqslant1,0\leqslant\bar{\phi}(x)\leqslant1,0\leqslant\phi(x)\leqslant1$，则有下列结论：

(1) 若

$$\int_S\phi f_i\mathrm{d}x=\int_S\phi^* f_i\mathrm{d}x\quad(i=1,2)\tag{3.13}$$

且

$$\phi^*(x)=\begin{cases}1, & \text{当 } f_0(x)>\sum_1^2 K_i f_i(x)\\ 0, & \text{当 } f_0(x)<\sum_1^2 K_i f_i(x)\end{cases}$$

这里 K_1,K_2 是常数，则

$$\int_S\phi^* f_0\mathrm{d}x\geqslant\int_S\phi f_0\mathrm{d}x\tag{3.14}$$

(2) 若 $\int_S \phi f_i \mathrm{d}x = \int_S \bar{\phi} f_i \mathrm{d}x \quad (i=1,2)$

$$\bar{\phi}(x) = \begin{cases} 1, & \text{当 } f_0(x) < \sum_1^2 K_i f_i(x) \\ 0, & \text{当 } f_0(x) > \sum_1^2 K_i f_i(x) \end{cases}$$

这里 K_1, K_2 是常数,则

$$\int_S \bar{\phi} f_0 \mathrm{d}x \leqslant \int_S \phi f_0 \mathrm{d}x \tag{3.15}$$

(3) 若(3.13)中"="号换成"≤"号,又 $K_1 \geqslant 0, K_2 \geqslant 0$,则(3.14)仍然成立.

证明:不难看出

$$(\phi^* - \phi)\left(f_0 - \sum_1^2 K_i f_i\right) \geqslant 0$$

于是

$$\int_S (\phi^* - \phi)\left(f_0 - \sum_1^2 K_i f_i\right) \mathrm{d}x \geqslant 0$$

所以

$$\int_S (\phi^* - \phi) f_0 \mathrm{d}x \geqslant \sum_1^2 K_i \int_S (\phi^* - \phi) f_i \mathrm{d}x = 0 \tag{3.16}$$

从而

$$\int_S \phi^* f_0 \mathrm{d}x \geqslant \int_S \phi f_0 \mathrm{d}x$$

这就证明了(3.14). 从(3.16)看,若在(3.13)中"="换成"≤",又 $K_i \geqslant 0 (i=1,2)$,则(3.14)仍成立.

从 $(\bar{\phi} - \phi)\left(f_0 - \sum_1^2 K_i f_i\right) \leqslant 0$ 不难推出(3.15)成立.

证毕.

定理 3.2 的证明:我们首先指出,对一切 $\theta_3 < \theta_1$ 及 $\theta_3 > \theta_2$ 均有

$$P(\underset{\sim}{X} \in W_0 \mid \theta_3) \leqslant \alpha \tag{3.17}$$

为确定计,不妨设 $\theta_3 < \theta_1$(当 $\theta_3 > \theta_2$ 时可进行类似的讨论),记 $f_j(\underset{\sim}{x}) = \prod_{i=1}^n f(x_i, \theta_j)$. 我们指出,必存在 K_1, K_2 使得

$$W_0 \overset{\mathrm{d}}{=} \left\{ \underset{\sim}{x} : C_1 < \sum_1^n V(x_i) < C_2 \right\}$$

$$= \left\{ \underset{\sim}{x} : f_3(\underset{\sim}{x}) < \sum_1^2 K_i f_i(\underset{\sim}{x}) \right\}$$

实际上，

$$\left\{ \underset{\sim}{x} : f_3(\underset{\sim}{x}) < \sum_1^2 K_i f_i(\underset{\sim}{x}) \right\}$$

$$= \left\{ \underset{\sim}{x} : S^n(\theta_3) \mathrm{e}^{Q(\theta_3) \sum_1^n V(x_i)} < \sum_1^2 K_i S^n(\theta_i) \mathrm{e}^{Q(\theta_i) \sum_1^n V(x_i)} \right\}$$

$$= \left\{ \underset{\sim}{x} : 1 < \sum_{j=1}^2 a_j \mathrm{e}^{b_j t} \right\} \quad (\underset{\sim}{x} = (x_1, \cdots, x_n))$$

这里

$$t = \sum_1^n V(x_i)$$

$$b_j = Q(\theta_j) - Q(\theta_3) > 0$$

$$a_j = K_j \left[\frac{S(\theta_j)}{S(\theta_3)} \right]^n \quad (j=1,2) \tag{3.18}$$

我们指出，对 C_1, C_2 必有 $a_1 > 0, a_2 < 0$ 使得

$$\left\{ t : 1 < \sum_1^2 a_j \mathrm{e}^{b_j t} \right\} = (C_1, C_2)$$

实际上，解方程组（a_1, a_2 是未知数）：

$$1 = \sum_1^2 a_j \mathrm{e}^{b_j C_i} \quad (i=1,2)$$

行列式

$$\Delta = \begin{vmatrix} \mathrm{e}^{b_1 C_1} & \mathrm{e}^{b_2 C_1} \\ \mathrm{e}^{b_1 C_2} & \mathrm{e}^{b_2 C_2} \end{vmatrix}$$

$$= \mathrm{e}^{b_1 C_2 + b_2 C_1} \left[\mathrm{e}^{(b_2 - b_1)(C_2 - C_1)} - 1 \right] > 0$$

于是有解

$$a_1 = \frac{1}{\Delta} (\mathrm{e}^{b_2 C_2} - \mathrm{e}^{b_2 C_1}) > 0$$

$$a_2 = \frac{1}{\Delta} (\mathrm{e}^{b_1 C_1} - \mathrm{e}^{b_1 C_2}) < 0$$

不难看出

$$\left\{t:1<\sum_1^2 a_j e^{b_j t}\right\}=\{t: e^{-b_1 t}-a_2 e^{(b_2-b_1)t}<a_1\}$$

但函数 $e^{-b_1 t}-a_2 e^{(b_2-b_1)t}$ 是 t 的严格凸函数，故

$$\left\{t:1<\sum_1^2 a_j e^{b_j t}\right\}=(C_1, C_2)$$

由(3.18)可求出 K_1, K_2，于是对此 K_1, K_2，有

$$\left\{\underset{\sim}{x}: f_3(\underset{\sim}{x})<\sum_1^2 K_j f_j(\underset{\sim}{x})\right\}$$

$$=\left\{\underset{\sim}{x}: C_1<\sum_1^n V(x_i)<C_2\right\}=W_0$$

令

$$\tilde{\phi}(\underset{\sim}{x})=\begin{cases}1, & \text{当 } \underset{\sim}{x}\in W_0\\ 0, & \text{当 } \underset{\sim}{x}\overline{\in} W_0\end{cases}$$

$$\phi(\underset{\sim}{x})\equiv\alpha$$

注意

$$\int\phi(\underset{\sim}{x})f_i(\underset{\sim}{x})\mathrm{d}\underset{\sim}{x}=\alpha=\int\tilde{\phi}(\underset{\sim}{x})f_i(\underset{\sim}{x})\mathrm{d}\underset{\sim}{x}$$

$(i=1,2)$. 从引理 3.1 知

$$P(\underset{\sim}{X}\in W_0\mid\theta_3)=\int\tilde{\phi}(\underset{\sim}{x})f_3(\underset{\sim}{x})\mathrm{d}\underset{\sim}{x}$$

$$\leqslant\int\phi(\underset{\sim}{x})f_3(\underset{\sim}{x})\mathrm{d}\underset{\sim}{x}=\alpha$$

这就证明了(3.17).

下面证明，对任何否定域 W，只要适合 $P(\underset{\sim}{X}\in W\mid\theta)\leqslant\alpha$（一切 $\theta\overline{\in}(\theta_1,\theta_2)$），则一定有

$$P(\underset{\sim}{X}\in W_0\mid\theta)\geqslant P(\underset{\sim}{X}\in W\mid\theta)\quad(\theta\in(\theta_1,\theta_2))\tag{3.19}$$

任意固定 $\theta_4\in(\theta_1,\theta_2)$，仍记

$$f_j(\underset{\sim}{x})=\prod_{i=1}^n f(x_i,\theta_j)\quad(j=1,2,3,4)$$

我们指出，必有 $K_1>0, K_2>0$ 使得

$$W_0=\left\{\underset{\sim}{x}: f_4(\underset{\sim}{x})>\sum_{j=1}^2 K_j f_j(\underset{\sim}{x})\right\}\tag{3.20}$$

实际上

$$\left\{\underset{\sim}{x}: f_4(\underset{\sim}{x})>\sum_{j=1}^2 K_j f_j(\underset{\sim}{x})\right\}$$

$$= \left\{ \underset{\sim}{x} : 1 > \sum_1^2 K_j \left(\frac{S(\theta_j)}{S(\theta_4)} \right)^n \exp \left\{ [Q(\theta_j) - Q(\theta_4)] \sum_1^n V(x_j) \right\} \right\}$$

$$= \left\{ \underset{\sim}{x} : 1 > \sum_1^2 a_j e^{b_j t} \right\}$$

这里

$$b_1 = Q(\theta_1) - Q(\theta_4) < 0$$

$$b_2 = Q(\theta_2) - Q(\theta_4) > 0$$

$$a_j = K_j \frac{S^n(\theta_j)}{S^n(\theta_4)}, t = \sum_1^n V(x_i)$$

与前面的论证类似,知可取 $a_1 > 0, a_2 > 0$ 使得

$$1 = \sum_1^2 a_j e^{b_j C_1} = \sum_1^2 a_j e^{b_j C_2}$$

因 $\sum_1^2 a_j e^{b_j t}$ 是 t 的严格凸函数,故

$$\left\{ t : \sum_1^2 a_j e^{b_j t} < 1 \right\} = (C_1, C_2)$$

从 $a_1 > 0, a_2 > 0$ 知有 $K_1 > 0, K_2 > 0$ 使得(3.20)成立.

取 $\phi^* = I_{W_0}, \phi = I_W$(示性函数),则

$$\int \phi^* f_i d\underset{\sim}{x} = P(\underset{\sim}{X} \in W_0 \mid \theta_i) = \alpha$$

$$\geqslant P(\underset{\sim}{X} \in W \mid \theta_i) = \int \phi f_i d\underset{\sim}{x} \quad (i = 1, 2)$$

根据引理 3.1 知

$$\int \phi^* f_4 d\underset{\sim}{x} \geqslant \int \phi f_4 d\underset{\sim}{x}$$

即有 $P(\underset{\sim}{X} \in W_0 \mid \theta_4) \geqslant P(\underset{\sim}{X} \in W \mid \theta_4)$,这就证明了(3.19).

证毕.

定理 3.3 的证明:为了证 W_0 的水平是 α,只须证明对一切 $\theta_3 \in (\theta_1, \theta_2)$ 有

$$P(\underset{\sim}{X} \in W_0 \mid \theta_3) \leqslant \alpha \tag{3.21}$$

记

$$f_j(\underset{\sim}{x}) = \prod_{i=1}^n f(x_i, \theta_j) \quad (j = 1, 2, 3)$$

我们指出存在 K_1, K_2 使得

$$W_0=\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})<\sum_1^2 K_j f_j(\underset{\sim}{x})\right\} \tag{3.22}$$

实际上

$$\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})<\sum_1^2 K_j f_j(\underset{\sim}{x})\right\}=\left\{\underset{\sim}{x}:1<\sum_1^2 a_j \mathrm{e}^{b_j t}\right\}$$

这里

$$b_1=Q(\theta_1)-Q(\theta_3)<0$$

$$b_2=Q(\theta_2)-Q(\theta_3)>0$$

$$a_j=K_j\left[\frac{S(\theta_j)}{S(\theta_3)}\right]^n \quad (j=1,2)$$

$$t=\sum_1^n V(x_i)$$

解方程组

$$1=\sum_1^2 a_j \mathrm{e}^{b_j C_1}=\sum_1^2 a_j \mathrm{e}^{b_j C_2}$$

可求得 $a_1>0,a_2>0$,从而可确定出 $K_1>0,K_2>0$. 由于 $\sum_1^2 a_j \mathrm{e}^{b_j t}$ 是 t 的严格凸函数,故

$$\left\{t:1<\sum_1^2 a_j \mathrm{e}^{b_j t}\right\}=(-\infty,C_1)\cup(C_2,\infty)$$

所以(3.22)成立.

取 $\phi(\underset{\sim}{x})\equiv\alpha \quad \bar{\phi}(\underset{\sim}{x})=I_{W_0}(\underset{\sim}{x})$,从引理 3.1 知

$$\int\bar{\phi}f_3\,\mathrm{d}\underset{\sim}{x}\leqslant\int\phi f_3\,\mathrm{d}\underset{\sim}{x}$$

即有 $P(\underset{\sim}{X}\in W_0\mid\theta_3)\leqslant\alpha$,这就证明了(3.21).

为了证明 W_0 的无偏性,我们首先指出对一切 $\theta_3\overline{\in}[\theta_1,\theta_2]$,必存在 K_1,K_2 使得

$$W_0=\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})>\sum_1^2 K_j f_j(\underset{\sim}{x})\right\} \tag{3.23}$$

不妨设 $\theta_3>\theta_2$(当 $\theta_3<\theta_1$ 时可作类似的论证),显然

$$\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})>\sum_1^2 K_j f_j(\underset{\sim}{x})\right\}=\left\{\underset{\sim}{x}:1>\sum_1^2 a_j \mathrm{e}^{b_j t}\right\}$$

这里

$$b_1 \overset{\mathrm{d}}{=} Q(\theta_1) - Q(\theta_3) < b_2 \overset{\mathrm{d}}{=} Q(\theta_2) - Q(\theta_3) < 0$$

$$a_j = K_j \left[\frac{S(\theta_j)}{S(\theta_3)} \right]^n \quad (j=1,2)$$

$$t = \sum_1^n V(x_i)$$

解方程组

$$1 = \sum_1^2 a_j \mathrm{e}^{b_j C_1} = \sum_1^2 a_j \mathrm{e}^{b_j C_2}$$

可求出 a_1, a_2 且 $a_1 < 0, a_2 > 0$,从而可求出 K_1, K_2.

注意

$$\left\{ t: 1 > \sum_1^2 a_j \mathrm{e}^{b_j t} \right\} = \{ t: a_2 < \mathrm{e}^{-b_2 t} - a_1 \mathrm{e}^{(b_1 - b_2) t} \}$$

但 $\mathrm{e}^{-b_2 t} - a_1 \mathrm{e}^{(b_1 - b_2) t}$ 是 t 的严格凸函数,故

$$\left\{ t: 1 > \sum_1^2 a_j \mathrm{e}^{b_j t} \right\} = (-\infty, C_1) \cup (C_2, \infty)$$

所以(3.23)成立.

令 $\phi^* = I_W(\underset{\sim}{x})$, $\phi \equiv \alpha$,则

$$\int \phi^* f_i \mathrm{d}\underset{\sim}{x} = P(\underset{\sim}{X} \in W_0 | \theta_i) = \alpha = \int \phi f_i \mathrm{d}\underset{\sim}{x} \quad (i=1,2)$$

从(3.23)和引理 3.1 知

$$\int \phi^* f_3 \mathrm{d}\underset{\sim}{x} \geqslant \int \phi f_3 \mathrm{d}\underset{\sim}{x}$$

即有 $P(\underset{\sim}{X} \in W_0 | \theta_3) \geqslant \alpha$. 这就证明了 W_0 是水平为 α 的无偏否定域.

设 W 是水平为 α 的任何无偏否定域,我们来证明:对一切 $\theta_3 \overline{\in} [\theta_1, \theta_2]$ 有

$$P(\underset{\sim}{X} \in W_0 \mid \theta_3) \geqslant P(\underset{\sim}{X} \in W \mid \theta_3) \tag{3.24}$$

实际上 $P(\underset{\sim}{X} \in W | \theta)$

$$= \int_W S^n(\theta) \prod_{i=1}^n h(x_i) \exp\left\{ Q(\theta) \sum_1^n V(x_i) \right\} \mathrm{d}\underset{\sim}{x}$$

既然假设 $Q(\theta)$ 是连续函数,易知 $P(\underset{\sim}{X} \in W | \theta)$ 也是 θ 的连续函数,根据无偏

性，$P(\underset{\sim}{X}\in W|\theta)\geqslant\alpha$（一切 $\theta\in[\theta_1,\theta_2]$），利用连续性知

$$P(\underset{\sim}{X}\in W|\theta_i)\geqslant\alpha\quad(i=1,2)$$

另一方面，W 的水平是 α，从而

$$P(\underset{\sim}{X}\in W|\theta_i)=\alpha\quad(i=1,2)$$

仍记 $\phi^*=I_{W_0}$，令 $\tilde{\phi}=I_W$，则

$$\int\phi^* f_i\mathrm{d}\underset{\sim}{x}=P(\underset{\sim}{x}\in W_0|\theta_i)=\alpha$$

$$=P(\underset{\sim}{X}\in W|\theta_i)=\int\tilde{\phi}f_i\mathrm{d}\underset{\sim}{x}\quad(i=1,2)$$

利用(3.23)和引理3.1知

$$\int\phi^* f_3\mathrm{d}\underset{\sim}{x}\geqslant\int\tilde{\phi}f_3\mathrm{d}\underset{\sim}{x}$$

即

$$P(\underset{\sim}{X}\in W_0|\theta_3)\geqslant P(\underset{\sim}{X}\in W|\theta_3)$$

故(3.24)成立. 这就证明了 W_0 是一致最大功效无偏否定域.

证毕.

定理3.4的证明：任给定 $\theta_1\neq\theta_0$，令

$$f_1(\underset{\sim}{x})=L(\underset{\sim}{x},\theta_0)=\prod_1^n f(x_i,\theta_0)$$

$$f_2(\underset{\sim}{x})=L(\underset{\sim}{x},\theta_0)t(\underset{\sim}{x})$$

$$f_3(\underset{\sim}{x})=L(\underset{\sim}{x},\theta_1)$$

我们指出，必存在 K_1,K_2 使得

$$W_0=\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})>\sum_1^2 K_j f_j(\underset{\sim}{x})\right\}\tag{3.25}$$

实际上

$$\left\{\underset{\sim}{x}:f_3(\underset{\sim}{x})>\sum_1^2 K_j f_j(\underset{\sim}{x})\right\}=\{\underset{\sim}{x}:\mathrm{e}^{bt(\underset{\sim}{x})}-a_1-a_2t(\underset{\sim}{x})>0\}$$

这里

$$b=Q(\theta_1)-Q(\theta_0)$$

$$a_j=K_j\left[\frac{S(\theta_0)}{S(\theta_1)}\right]^n\quad(j=1,2)$$

解方程

$$e^{bC_1} - a_1 - a_2 C_1 = e^{bC_2} - a_1 - a_2 C_2 = 0$$

对给定的 C_1, C_2，可求出 a_1, a_2，从而可确定 K_1, K_2，由于 $e^{bt} - a_1 - a_2 t$ 是 t 的严格凸函数，知

$$\{t: e^{bt} - a_1 - a_2 t > 0\} = (-\infty, C_1) \cup (C_2, \infty)$$

所以

$$\left\{\underset{\sim}{x}: f_3(\underset{\sim}{x}) > \sum_1^2 K_j f_j(\underset{\sim}{x})\right\} = \{\underset{\sim}{x}, t(\underset{\sim}{x}) < C_1, \text{或} > C_2\} = W_0$$

即(3.25)成立.

为了证明 W_0 的无偏性，我们来证：

$$P(\underset{\sim}{X} \in W_0 \mid \theta_1) \geqslant \alpha \quad (\theta_1 \neq \theta_0) \tag{3.26}$$

令 $\phi^* = I_{W_0}(\underset{\sim}{x})$，$\phi \equiv \alpha$，易知

$$\int \phi^* f_1 \mathrm{d}\underset{\sim}{x} = P(\underset{\sim}{X} \in W_0 \mid \theta_0) = \alpha = \int \phi f_1 \mathrm{d}\underset{\sim}{x}$$

$$\int \phi^* f_2 \mathrm{d}\underset{\sim}{x} = E_{\theta_0}[I_{W_0}(\underset{\sim}{X}) t(\underset{\sim}{X})] = \alpha E_{\theta_0} t(\underset{\sim}{X}) = \int \phi f_2 \mathrm{d}\underset{\sim}{x}$$

(这里用了条件(3.10))，利用(3.25)和引理 3.1 知

$$\int \phi^* f_3 \mathrm{d}\underset{\sim}{x} \geqslant \int \phi f_3 \mathrm{d}\underset{\sim}{x}$$

即 $P(\underset{\sim}{X} \in W_0 \mid \theta_1) \geqslant \alpha$，这就证明了(3.26).

设 W 是水平为 α 的任一无偏否定域，我们来证：对一切 $\theta_1 \neq \theta_0$ 有

$$P(\underset{\sim}{X} \in W_0 \mid \theta_1) \geqslant P(\underset{\sim}{X} \in W \mid \theta_1) \tag{3.27}$$

为此，我们首先指出：

$$E_{\theta_0}(I_W(\underset{\sim}{X}) t(\underset{\sim}{X})) = \alpha E_{\theta_0} t(\underset{\sim}{X}) \tag{3.28}$$

令

$$\psi(\theta) = \int_W \prod_1^n f(x_i, \theta) \mathrm{d}\underset{\sim}{x}$$

易知 $\psi(\theta)$ 是 θ 的可微函数. 注意

$$\psi(\theta_0) = P(\underset{\sim}{X} \in W \mid \theta_0) = \alpha$$

根据 W 的无偏性 $\psi(\theta) \geqslant \alpha = \psi(\theta_0)$（一切 θ），所以 $\psi(\theta)$ 在 θ_0 达最小值，从而 $\psi'(\theta_0) = 0$.

但是

$$\psi(\theta)=C(\theta)\int_W \prod_1^n h(x_j)\mathrm{e}^{Q(\theta)t(\underset{\sim}{x})}\mathrm{d}\underset{\sim}{x}$$

$$(C(\theta)=[S(\theta)]^n)$$

于是

$$\psi'(\theta)=C'(\theta)\int_W \prod_1^n h(x_j)\mathrm{e}^{Q(\theta)t(\underset{\sim}{x})}\mathrm{d}\underset{\sim}{x}+$$

$$C(\theta)\int_W \prod_1^n h(x_j)\mathrm{e}^{Q(\theta)t(\underset{\sim}{x})}t(\underset{\sim}{x})Q'(\theta)\mathrm{d}\underset{\sim}{x}$$

$$=\frac{C'(\theta)}{C(\theta)}E_\theta I_W(\underset{\sim}{X})+Q'(\theta)E_\theta[I_W(\underset{\sim}{X})t(\underset{\sim}{X})]$$

但

$$1\equiv\int C(\theta)\prod_1^n h(x_j)\mathrm{e}^{Q(\theta)t(\underset{\sim}{x})}\mathrm{d}\underset{\sim}{x}$$

两边微商知

$$0=\frac{C'(\theta)}{C(\theta)}+Q'(\theta)E_\theta t(\underset{\sim}{X})$$

从 $\psi'(\theta_0)=0$,推知

$$-Q'(\theta_0)E_{\theta_0}t\cdot\alpha+Q'(\theta_0)E_{\theta_0}[I_W\cdot t]=0$$

于是(3.28)成立.

仍令 $\phi^*=I_{W_0}$,又令 $\tilde{\phi}=I_W$,则

$$\int\phi^* f_1\mathrm{d}\underset{\sim}{x}=P(\underset{\sim}{X}\in W_0|\theta_0)=\alpha=P(\underset{\sim}{X}\in W|\theta_0)=\int\tilde{\phi}f_1\mathrm{d}\underset{\sim}{x}$$

从(3.28)知

$$\int\phi^* f_2\mathrm{d}\underset{\sim}{x}=E_{\theta_0}[I_{W_0}(\underset{\sim}{X})t(\underset{\sim}{X})]$$

$$=\alpha E_{\theta_0}t(\underset{\sim}{X})=E_{\theta_0}[I_{W_0}(\underset{\sim}{X})t(\underset{\sim}{X})]$$

$$=\int\tilde{\phi}f_2\mathrm{d}\underset{\sim}{x}$$

利用(3.25)及引理 3.1 知

$$\int\phi^* f_3\mathrm{d}\underset{\sim}{x}\geqslant\int\tilde{\phi}f_3\mathrm{d}\underset{\sim}{x}$$

即

$$P(\underset{\sim}{X}\in W_0|\theta_1)\geqslant P(\underset{\sim}{X}\in W|\theta_1)$$

这就证明了(3.27).

证毕.

§4 广义似然比检验

上一节处理问题虽很圆满，但所讨论的情况太特殊，对总体的概率分布有很强的限制，许多常见的检验问题就对付不了. 本节介绍一种检验法，它适用的范围极广，当样本量 n 较大时，否定域往往相当好(即第二类错误的概率比较小)，虽然它不一定是一致最大功效的. 多数情况下可以证明 $\lim\limits_{n} \rho_{W_0}(\theta)=1$(对一切 $\theta\in\Theta_1$)，这里 $H_a:\theta\in\Theta_1$ 是备择假设，$\rho_{W_0}(\theta)$ 是参数真值是 θ 时检验法 W_0 拒绝零假设的概率.

设 X 的分布密度是 $f(x,\theta)$，$\theta=(\theta_1,\cdots,\theta_m)\in\Theta\subset R^m$(对离散型分布可进行类似的讨论).

设 Θ_0 是 Θ 的非空真子集，研究检验问题：

$$H_0:\theta\in\Theta_0\leftrightarrow H_a:\theta\in\Theta-\Theta_0$$

设 $\underset{\sim}{x}=(x_1,\cdots,x_n)$ 是样本 $\underset{\sim}{X}=(X_1,\cdots,X_n)$ 的值. 似然函数 $L(\underset{\sim}{x},\theta)=\prod\limits_{1}^{n} f(x_i,\theta)$，令

$$L(\hat{\Theta}_0)\overset{\mathrm{d}}{=}\sup_{\theta\in\Theta_0} L(\underset{\sim}{x},\theta)$$

$$L(\hat{\Theta})\overset{\mathrm{d}}{=}\sup_{\theta\in\Theta} L(\underset{\sim}{x},\theta)$$

定义 4.1 称

$$\lambda(\underset{\sim}{x})\overset{\mathrm{d}}{=}\frac{L(\hat{\Theta})}{L(\hat{\Theta}_0)}$$

为样本值 $\underset{\sim}{x}=(x_1,\cdots,x_n)$ 的广义似然比.

显然 $\lambda(\underset{\sim}{x})\geqslant 1$，直观上看，若 H_0 是真的，则 θ 的最大似然估计应该很大可能属于 Θ_0，从而 $\lambda(\underset{\sim}{x})$ 应该接近于 1；反之，若 $\lambda(\underset{\sim}{x})$ 的值

太大就应该否定 H_0，这样应取否定域

$$W_0=\{\underset{\sim}{x}:\lambda(\underset{\sim}{x})>\lambda_0\} \tag{4.1}$$

其中 λ_0 满足

$$\sup_{\theta\in\Theta_0} P(\underset{\sim}{X}\in W_0\,|\,\theta)=\alpha \tag{4.2}$$

这里 α 是预先给定的检验水平($0<\alpha<1$).

这样的检验法叫做广义似然比检验法，简称似然比检验法.这个方法在许多情形下常可导出有实用价值的具体否定域.在使用广义似然比检验法时，关键在于能否找出 λ_0 满足(4.2).

我们指出，广义似然比 λ 是充分统计量的函数，从而否定域(4.1)常常是由充分统计量来确定. 实际上，若 $\varphi(\underset{\sim}{X})$ 是充分统计量(可以是向量)，则有 $L(\underset{\sim}{x},\theta)=g[\varphi(\underset{\sim}{x}),\theta]\cdot h(\underset{\sim}{x})$，从而

$$\lambda(\underset{\sim}{x})=\frac{\sup\limits_{\theta\in\Theta} g[\varphi(\underset{\sim}{x}),\theta]}{\sup\limits_{\theta\in\Theta_0} g[\varphi(\underset{\sim}{x}),\theta]}=l[\varphi(\underset{\sim}{x})]$$

所以否定域

$$W_0=\{\underset{\sim}{x}:\lambda(\underset{\sim}{x})>\lambda_0\}=\{\underset{\sim}{x}:\varphi(\underset{\sim}{x})\in B\}$$

这里 B 是适当的集合.

由于正态总体在实际工作中最常见，数学处理上也比较成熟，下面讨论正态总体的各种检验问题.

(1) 设 $X\sim N(\mu,\sigma^2)$，σ^2 已知，$\mu\in\Theta=(-\infty,\infty)$ 考虑检验问题：

$$H_0:\mu=\mu_0\leftrightarrow H_a:\mu\neq\mu_0$$

给定 $\alpha=0.05$，样本是 $X_1,\cdots,X_n$，如何检验 H_0？

这个检验问题已在§3中解决了，现在用广义似然比检验法来试一试.

似然函数

$$L(\underset{\sim}{x},\mu)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_1^n(x_i-\mu)^2\right)$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_1^n (x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right]\right)$$

因为 $\Theta_0=\{\mu_0\}$，易知 $L(\hat{\Theta}_0)=L(\underset{\sim}{x},\mu_0)$，

$$L(\hat{\Theta})=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_1^n (x_i-\bar{x})^2\right),$$

于是广义似然比 $\lambda=e^{\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2}$，可见否定域

$$W_0=\{\underset{\sim}{x}:\lambda(\underset{\sim}{x})>\lambda_0\}=\{\underset{\sim}{x}:|\bar{x}-\mu_0|>C\}$$

其中 C 满足

$$P(|\bar{X}-\mu_0|>C|\mu_0)=0.05$$

因为 $\mu=\mu_0$ 时，$\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\sim N(0,1)$，故利用正态分布表知 $C=\frac{1.96}{\sqrt{n}}\sigma$. 这样得到的否定域与§3 中获得的否定域是一致的，从§3 中知它是一致最大功效无偏的.

(2) 设 $X\sim N(\mu,\sigma^2)$，$\Theta=\{(\mu,\sigma^2):-\infty<\mu<\infty,0<\sigma^2<\infty\}$，方差 σ^2 未知，检验问题是：

$$H_0:\mu=\mu_0\leftrightarrow H_a:\mu\neq\mu_0$$

这类问题在实际工作中更常见，§1 中例 1.2 即是一例.

我们用广义似然比检验法导出合适的否定域. 记 $\Theta_0=\{(\mu_0,\sigma^2):\sigma^2>0\}$，设 $\underset{\sim}{x}=(x_1,\cdots,x_n)$ 是样本 $\underset{\sim}{X}=(X_1,\cdots,X_n)$ 的值. $\underset{\sim}{x}$ 的似然函数

$$L(\underset{\sim}{x},\theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n (x_i-\mu)^2\right\}$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_1^n (x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right]\right\}$$

$(\theta\overset{\mathrm{d}}{=\!=}(\mu,\sigma^2))$. 易知

$$L(\hat{\Theta})=\sup_\theta L(\underset{\sim}{x},\theta)=\left[\frac{n}{2\pi\sum_1^n (x_i-\bar{x})^2}\right]^{\frac{n}{2}} e^{-\frac{n}{2}}$$

$$L(\hat{\Theta}_0)=\sup_{\theta\in\Theta_0} L(\underset{\sim}{x},\theta) = \left[\frac{n}{2\pi\sum_1^n (x_i-\mu_0)^2}\right]^{\frac{n}{2}} e^{-\frac{n}{2}}$$

于是广义似然比

$$\lambda=\left[\frac{\sum_1^n (x_i-\mu_0)^2}{\sum_1^n (x_i-\bar{x})^2}\right]^{\frac{n}{2}}=\left(1+\frac{T^2}{n-1}\right)^{\frac{n}{2}}$$

其中

$$T=\frac{\sqrt{n(n-1)}(\bar{x}-\mu_0)}{\sqrt{\sum_1^n (x_i-\bar{x})^2}} \tag{4.3}$$

显然 λ 是 T^2 的严格增函数,故广义似然比检验的否定域为:

$$W_0=\{(x_1,\cdots,x_n):|T|>C\} \tag{4.4}$$

其中 C 适合

$$P(\underset{\sim}{X}\in W_0|(\mu_0,\sigma))=\alpha \quad (一切\ \sigma)$$

从第二章知 $\mu=\mu_0$ 时,统计量 T 服从 $n-1$ 个自由度的 t 分布,故(4.4)中的 C 可从 t 分布表中查出. 例如 $n=10$ 时,自由度为 9, $\alpha=0.05$时查出 $C=2.262$.

否定域(4.4)有何优良性呢? 可以证明它是一致最大功效无偏的否定域. 证明过程较复杂,超出了本书范围. 读者如有兴趣,可参看文献[12].

例 4.1 用某仪器间接测量某物质的温度(℃),测 5 次,数据是 1 250,1 265,1 245,1 260,1 275,而用别的精确方法得到温度的值为 1 277(可看作真值). 问该仪器测温有无系统偏差?

由于测量值一般服从正态分布 $N(\mu,\sigma^2)$,这个问题就是检验 $H_0:\mu=1\,277$,但是方差是未知的,根据上面讲的,应取否定域

$$W_0=\{(x_1,\cdots,x_5):|T|>C\}$$

其中 T 由(4.3)给出($n=5$),查 t 分布表知 $\alpha=0.05$ 时,$C=$

2.776. 故

$$W_0=\{(x_1,\cdots,x_5):|T|>2.776\}$$

根据所给的样本值，计算得 $|T|=\dfrac{18}{5.339}>2.776$，故应否定 H_0. 换句话说，该仪器测温有系统偏差.

(3) 设 $X\sim N(\mu,\sigma^2)$，σ^2 未知，$x_1,\cdots,x_n$ 是样本值，对于检验问题：

$$H_0:\mu\leqslant\mu_0\leftrightarrow H_a:\mu>\mu_0$$

可用广义似然比检验法导出否定域：

$$W_0=\{(x_1,\cdots,x_n):T>C_1\} \tag{4.5}$$

其中 T 的表达式见(4.3)，C_1 由方程

$$\int_{C_1}^{\infty}t_{n-1}(y)\mathrm{d}y=\alpha$$

确定. 这里 $t_{n-1}(y)$是 $n-1$ 个自由度的 t 分布的分布密度. C_1 可查表得到.

否定域(4.5)在直观上甚为合理. 数学上可以证明，它是一致最大功效无偏的，参见文献[12].

类似地，对于检验问题：

$$H_0:\mu\geqslant\mu_0\leftrightarrow H_a:\mu<\mu_0$$

可用广义似然比检验法导出否定域：

$$W_0=\{(x_1,\cdots,x_n):T<C_2\} \tag{4.6}$$

其中 T 的表达式参见(4.3)，C_2 满足方程

$$\int_{C_2}^{\infty}t_{n-1}(y)\mathrm{d}y=1-\alpha$$

C_2 可查 t 分布表得到. 可以证明否定域(4.6)也是一致最大功效无偏的.

(4) 设 $X\sim N(\mu,\sigma^2)$，μ 未知，考虑检验问题：

$$H_0:\sigma^2=\sigma_0^2\leftrightarrow H_a:\sigma^2\neq\sigma_0^2$$

给定 $\alpha\in(0,1)$，设样本 $\underset{\sim}{X}=(X_1,\cdots,X_n)$的值是 $\underset{\sim}{x}=(x_1,\cdots,$

x_n). 利用广义似然比检验法可导出否定域为

$$W_0=\{(x_1,\cdots,x_n):u>C_2 \text{ 或 } <C_1\} \tag{4.7}$$

其中

$$u=\frac{1}{\sigma_0^2}\sum_1^n(x_i-\overline{x})^2 \tag{4.8}$$

$C_1<C_2$ 满足:

$$P(\underset{\sim}{X}\in W_0\,|\,\mu,\sigma_0)=\alpha \tag{4.9}$$

可以证明,在 H_0 成立的条件下 $u=u(X_1,\cdots,X_n)$服从 $n-1$ 个自由度的χ^2分布. (4.9)等价于

$$\int_0^{C_1}g_{n-1}(y)\mathrm{d}y+\int_{C_2}^{\infty}g_{n-1}(y)\mathrm{d}y=\alpha \tag{4.10}$$

这里 $g_n(y)$是 n 个自由度的χ^2分布的分布密度.

否定域(4.7)在直观上甚为合理. 数学上可以证明,如果 C_1,C_2 满足(4.10)及

$$\int_{C_1}^{C_2}g_{n+1}(y)\mathrm{d}y=1-\alpha$$

则否定域(4.7)是一致最大功效无偏的.

在实用上常取 $C_1<C_2$,满足

$$\left.\begin{aligned}\int_0^{C_1}g_{n-1}(y)\mathrm{d}y=\frac{\alpha}{2}\\ \int_{C_2}^{\infty}g_{n-1}(y)\mathrm{d}y=\frac{\alpha}{2}\end{aligned}\right. \tag{4.11}$$

这样的 C_1,C_2 可通过查χ^2分布表很易得到.

例 4.2 某车间生产铜丝,生产一向比较稳定. 今从产品中随便抽取 10 根检查其折断力,得数据如下:

578,512,570,568,572,570,570,572,596,654

问:是否可相信该车间的铜丝的折断力的方差为 64?

用 X 表示铜丝的折断力,可以认为 $X\sim N(\mu,\sigma^2)$. 我们的任务是根据样本值 $x_1,\cdots,x_{10}$来检验假设 $H_0:\sigma^2=64$.

根据上面所讲的，应取否定域

$$W_0=\{(x_1,\cdots,x_{10}):u< C_1 \text{ 或 } > C_2\}$$

其中 $u=\frac{1}{64}\sum_{1}^{10}(x_i-\overline{x})^2$.

给定 $\alpha=0.05$，从χ^2分布表查得 $C_1=2.70$，$C_2=19.0$ 满足(4.11).

根据所给的样本值 $x_1=578,\cdots,x_{10}=584$，算得 $u=10.65$. 由于 $2.70<10.65<19.0$，故观察到的样本未落入否定域，从而应接受假设 $H_0:\sigma^2=64$.

(5) 设 $X\sim N(\mu,\sigma^2)$，研究检验问题：

$$H_0:\sigma^2\leqslant\sigma_0^2\leftrightarrow H_a:\sigma^2>\sigma_0^2$$

用广义似然比检验法导出的否定域是

$$W_0=\{(x_1,\cdots,x_n): u>C_2\} \tag{4.12}$$

其中 u 的表达式见(4.8)，C_2 满足：

$$\int_{C_2}^{\infty}g_{n-1}(y)\mathrm{d}y=\alpha$$

这里 $g_{n-1}(y)$是 $n-1$ 个自由度的χ^2分布的密度. 数学上可以证明否定域(4.12)是一致最大功效无偏的.

类似地，对于检验问题：

$$H_0:\sigma^2\geqslant\sigma_0^2\leftrightarrow H_a:\sigma^2<\sigma_0^2$$

用广义似然比检验法导出的否定域是

$$W_0=\{(x_1,\cdots,x_n):u<C_1\} \tag{4.13}$$

其中 u 的表达式见(4.8)，C_1 满足

$$\int_{0}^{C_1}g_{n-1}(y)\mathrm{d}y=\alpha$$

可以证明，否定域(4.13)是一致最大功效无偏的.

(6) 研究两个正态总体的假设检验问题. 设 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，$X$ 的样本是 $\underset{\sim}{X}=(X_1,\cdots,X_{n_1})'$，$Y$ 的样本是 $\underset{\sim}{Y}=(Y_1,\cdots,Y_{n_2})'$，我们设两个样本相互独立. 研究下列检验问题：

$$H_0:\sigma_1^2=\sigma_2^2\leftrightarrow H_a:\sigma_1^2\neq\sigma_2^2$$

给定 $\alpha\in(0,1)$，利用广义似然比检验法可导出否定域是

$$W_0=\{(x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}):F<C_1\text{ 或 }>C_2\} \tag{4.14}$$

其中

$$F=\frac{\sum_1^{n_1}(x_i-\overline{x})^2/(n_1-1)}{\sum_1^{n_2}(y_i-\overline{y})^2/(n_2-1)} \tag{4.15}$$

这里

$$\overline{x}=\frac{1}{n_1}\sum_1^{n_1}x_i,\quad \overline{y}=\frac{1}{n_2}\sum_1^{n_2}y_i$$

怎样确定 C_1,C_2 使得 W_0 的检验水平是给定的 α 呢？这就需要研究统计量 F(将 x_i,y_j 都看成是随机变量)的概率分布. 注意

$$\frac{1}{\sigma_1^2}\sum_1^{n_1}(X_i-\overline{X})^2\sim\chi^2(n_1-1)$$

$$\frac{1}{\sigma_2^2}\sum_1^{n_2}(Y_j-\overline{Y})^2\sim\chi^2(n_2-1)$$

而且二者相互独立，故在 H_0 下统计量 F 的分布是完全确定的，与 $\mu_1,\sigma_1,\mu_2,\sigma_2$ 无关. 当然 F 的分布与 n_1,n_2 有关，只要 F 的分布求出来了，则不难确定(4.14)中的 C_1,C_2.

定理 4.1 设 $\xi\sim\chi^2(n_1)$，$\eta\sim\chi^2(n_2)$ 且 ξ 与 η 相互独立($n_1\geqslant 1,n_2\geqslant 1$)，则 $\zeta\overset{\mathrm{d}}{=}\frac{\xi/n_1}{\eta/n_2}$的分布密度是

$$f_{n_1,n_2}(u)=\begin{cases}0, & u\leqslant 0\\ \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}}u^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}u\right)^{-\frac{n_1+n_2}{2}}, & u>0\end{cases} \tag{4.16}$$

这个结论在初等概率论中已证明过，证明方法是直接计算两个相互独立的随机变量的商的分布密度，无非是算一算积分，这里

从略(参看[1]). 表达式(4.16)以后并不用到,读者不必记住.

定义 4.2 称随机变量 X 服从自由度为 n_1,n_2 的 F 分布(简记为 $X\sim F(n_1,n_2)$),若它的分布密度有表达式(4.16),这时 n_1 叫做第一自由度,n_2 叫做第二自由度.

从(4.15)和定理 4.1 知统计量 F 服从自由度为 n_1-1,n_2-1 的 F 分布. F 分布的分位数表已经造出,查 F 分布表不难找到 $C_1<C_2$ 满足

$$\int_0^{C_1} f_{n_1-1,n_2-1}(u)\mathrm{d}u+\int_{C_2}^{\infty} f_{n_1-1,n_2-1}(u)\mathrm{d}u=\alpha \qquad (4.17)$$

于是在 H_0 下 $P(F<C_1$ 或 $>C_2)=\alpha$,从而否定域(4.14)的检验水平是 α.

数学上可以证明:如果 $C_1<C_2$ 满足(4.17)及下式

$$\int_{a_1}^{a_2} B_{\frac{1}{2}(n_1+1),\frac{1}{2}(n_2-1)}(y)\mathrm{d}y=1-\alpha$$

这里

$$a_i=\frac{n_1-1}{n_2-1}C_i\Big/\left(\frac{n_1-1}{n_2-1}C_i+1\right),\quad i=1,2$$

$B_{r,s}(y)$ 是 β 分布的密度函数,则否定域(4.14)是一致最大功效无偏的.

在实用上常取 C_1,C_2 满足

$$\int_0^{C_1} f_{n_1-1,n_2-1}(y)\mathrm{d}y=\int_{C_2}^{\infty} f_{n_1-1,n_2-1}(y)\mathrm{d}y=\frac{\alpha}{2} \qquad (4.18)$$

注意

$$\int_0^{C_1} f_{n_1-1,n_2-1}(y)\mathrm{d}y=P(F<C_1\mid H_0)=P\left(\frac{1}{F}>\frac{1}{C_1}\middle| H_0\right)$$

$$=\int_{C_1^{-1}}^{\infty} f_{n_2-1,n_1-1}(y)\mathrm{d}y$$

故 C_1 及 C_2 均可利用 F 分布表得到.

例 4.3(见例 1.3) 已知 70 ℃下的强力数据是 20.5,18.8,19.8,20.9,21.5,19.5,21.0,21.2;80 ℃下的强力数据是 17.7,

20.3,20.0,18.8,19.0,20.1,20.2,19.1. 问:70 ℃与 80 ℃下的强力的分散程度有无差别?

设 70 ℃下的强力为 X,80 ℃下的强力为 Y,实践经验表明,$X \sim N(\mu_1, \sigma_1^2)$,$Y \sim N(\mu_2, \sigma_2^2)$,所提问题化为检验假设 $H_0: \sigma_1^2 = \sigma_2^2$.

根据上面介绍的理论,应取统计量(4.15)及否定域(4.14),其中 C_1, C_2 由(4.18)确定. 本例中,$n_1 = n_2 = 8$,取 $\alpha = 0.05$,查表知 $C_1 = \dfrac{1}{4.99}$,$C_2 = 4.99$. 根据所给数据 $x_1 = 20.5, \cdots, x_8 = 21.2$,$y_1 = 17.7, \cdots, y_8 = 19.1$,算得 $F = 1.07$. 由于 $\dfrac{1}{4.99} < 1.07 < 4.99$,故样本未落入否定域,应接受 H_0,即认为 70 ℃与 80 ℃下的强力分散程度相同.

对于两个正态总体的检验问题:

$$H_0: \sigma_1^2 \leqslant \sigma_2^2 \leftrightarrow H_a: \sigma_1^2 > \sigma_2^2$$

应取否定域:

$$W_0 = \{F > C_2\} \tag{4.19}$$

这里 F 的表达式见(4.15),C_2 由下式确定:

$$\int_{C_2}^{\infty} f_{n_1-1, n_2-1}(y) \mathrm{d}y = \alpha$$

C_2 可由 F 分布表查出.

数学上可以证明,否定域(4.19)是一致最大功效无偏的.

类似地,对于检验问题:

$$H_0: \sigma_1^2 \geqslant \sigma_2^2 \leftrightarrow H_a: \sigma_1^2 < \sigma_2^2$$

应取否定域:

$$W_0 = \{F < C_1\} \tag{4.20}$$

其中 F 由(4.15)确定,C_1 由下式确定.

$$\int_0^{C_1} f_{n_1-1, n_2-1}(y) \mathrm{d}y = \alpha$$

由于$\int_{C_1^{-1}}^{\infty} f_{n_2-1,n_1-1}(y)\mathrm{d}y=\alpha$，可利用 F 分布表得到 C_1 的值.

数学上可以证明：否定域(4.20)是一致最大功效无偏的.

现在研究 $\sigma_1^2=\sigma_2^2$ 时的下列检验问题.

$$H_0: \mu_1=\mu_2 \leftrightarrow H_a: \mu_1 \neq \mu_2$$

利用广义似然比检验法可导出(推导从略)否定域为：

$$W_0=\{(x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}): |T|>C\} \tag{4.21}$$

其中

$$T=\frac{\overline{x}-\overline{y}}{\sqrt{\sum_1^{n_1}(x_i-\overline{x})^2+\sum_1^{n_2}(y_i-\overline{y})^2}} \cdot \sqrt{\frac{n_1 n_2(n_1+n_2-2)}{n_1+n_2}} \tag{4.22}$$

给定检验水平 α，怎样确定(4.21)中的 C 呢？这就需要研究 T(看成随机变量)的分布.

定理 4.2 设 $X_1,\cdots,X_{n_1}$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本，$Y_1,\cdots,Y_{n_2}$ 是来自 $N(\mu_2,\sigma_2^2)$ 的样本，且$(X_1,\cdots,X_{n_1})$与$(Y_1,\cdots,Y_{n_2})$相互独立，则在 $\sigma_1^2=\sigma_2^2$，$\mu_1=\mu_2$ 的条件下，统计量 T(见(4.22)，其中的 x_i,y_j 分别换成 X_i,Y_j)服从 n_1+n_2-2 个自由度的 t 分布.

证明：记 $\xi=\sum_1^{n_1}(X_i-\overline{X})^2$，$\eta=\sum_1^{n_2}(Y_i-\overline{Y})^2$，由于$(X_1,\cdots,X_{n_1})$与$(Y_1,\cdots,Y_{n_2})$相互独立，故$(\overline{X},\xi)$与$(\overline{Y},\eta)$相互独立. 根据第二章定理 3.3，$\overline{X}$ 与 ξ 独立，$\overline{Y}$ 与 η 独立，$\frac{1}{\sigma_1^2}\xi \sim \chi^2(n_1-1)$，$\frac{1}{\sigma_2^2}\eta \sim \chi^2(n_2-1)$，于是 $\overline{X},\xi,\overline{Y},\eta$ 相互独立. 从而 $\overline{X}-\overline{Y}$ 与 $\xi+\eta$ 相互独立. 当 $\mu_1=\mu_2$，$\sigma_1^2=\sigma_2^2=\sigma^2$ 时，$\overline{X}-\overline{Y} \sim N\left(0,\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot\sigma^2\right)$，$\frac{1}{\sigma^2}(\xi+\eta) \sim \chi^2(n_1+n_2-2)$，于是

$$T=\frac{(\overline{X}-\overline{Y})\Big/\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\sqrt{\frac{1}{(n_1+n_2-2)\sigma^2}(\xi+\eta)}}\sim t(n_1+n_2-2)$$

证毕.

根据定理 4.2,查 t 分布表有 C 满足

$$P(|T|>C|H_0)=\alpha$$

此时,否定域(4.21)的检验水平是 α.

否定域(4.21)在直观上甚为合理. 数学上可以证明,它是一致最大功效无偏的.

例 4.4(例 4.3 的继续) 问:70 ℃下的强力与 80 ℃下的强力有无差别?

从例 4.3 知两个温度下的强力的方差可以认为相等,即 $\sigma_1^2=\sigma_2^2$,故问题化为在 $\sigma_1^2=\sigma_2^2$ 下检验 $H_0:\mu_1=\mu_2$.

根据上面介绍的方法,应取否定域(4.21). 设 $\alpha=0.05$,现在 $n_1=n_2=8$,查 t 分布表(自由度为 $n_1+n_2-2=14$)得 $C=2.145$,根据所给的 16 个数据(见例 4.3),算得统计量 T 的数值为 2.161,因为 $T=2.161>2.145$,故应否定假设 H_0,换句话说,70 ℃下的强力与 80 ℃下的强力有差别,前者确实大些.

例 4.5 研究口服避孕药对妇女血压的影响. 对某公司工作的 35 岁至 39 岁的非怀孕妇女用抽查方法收集到下列数据. 有 8 人使用口服避孕药,8 人的血压(收缩压)的平均值是 132.86(单位是 mmHg 柱),标准差是 15.35;有 21 人未使用口服避孕药,血压的平均值是 127.44,标准差是 18.23. 问:这两种血压平均值的差异是否“显著”?

我们假定使用口服避孕药的妇女的血压(收缩压)服从正态分布 $N(\mu_1,\sigma_1^2)$,不使用口服避孕药的妇女的血压(收缩压)服从正态分布 $N(\mu_2,\sigma_2^2)$. 假定 $\sigma_1^2=\sigma_2^2$. 问题化为检验假设 $H_0:\mu_1=\mu_2$.

使用统计量(4.22),现在$\overline{x}=132.86$,$\overline{y}=127.44$,$n_1=8$,$n_2=$

21, $\sum_{1}^{n_1}(x_i-\overline{x})^2=7\times(15.35)^2$, $\sum_{1}^{n_2}(y_i-\overline{y})^2=20\times(18.23)^2$, 可算出统计量 T 的值为 0.74,设检验水平 $\alpha=0.05$,查 t 分布的临界值表知临界值 $\lambda=2.052$. 现在 $|T|=0.74<2.052$. 故不能拒绝 H_0. 即认为两个平均值无显著差异.

在 $\sigma_1^2=\sigma_2^2$ 的前提下,对于检验问题:

$$H_0:\mu_1\leqslant\mu_2\leftrightarrow H_a:\mu_1>\mu_2$$

应取否定域:

$$W_0=\{(x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}):T>C_2\} \tag{4.23}$$

其中 T 由(4.22)给出,C_2 满足

$$\int_{C_2}^{\infty}t_{n_1+n_2-2}(y)\mathrm{d}y=\alpha$$

这里 $t_n(y)$ 是 n 个自由度的 t 分布的密度.

可以证明,否定域(4.23)是一致最大功效无偏的.

类似地,对于检验问题:

$$H_0:\mu_1\geqslant\mu_2\leftrightarrow H_a:\mu_1<\mu_2$$

应取否定域:

$$W_0=\{(x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}):T<C_1\} \tag{4.24}$$

其中 T 由(4.22)确定,C_1 满足

$$\int_{-\infty}^{C_1}t_{n_1+n_2-2}(y)\mathrm{d}y=\alpha$$

也可以证明,否定域(4.24)是一致最大功效无偏的.

现在研究 $\sigma_1^2\neq\sigma_2^2$ 时的下列检验问题:

$$H_0:\mu_1=\mu_2\leftrightarrow H_a:\mu_1\neq\mu_2$$

这就是著名的 Behrens - Fisher 问题. 此时难以使用广义似然比检验法,我们采用下列解决方法. 设 $X_1,\cdots,X_{n_1}$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本,$Y_1,\cdots,Y_{n_2}$ 是来自 $N(\mu_2,\sigma_2^2)$的样本,二样本相互独立. 令

$$\overline{X}=\frac{1}{n_1}\sum_{1}^{n_1}X_i,\quad \overline{Y}=\frac{1}{n_2}\sum_{1}^{n_2}Y_i,$$

$$S_1^2=\frac{1}{n_1-1}\sum_1^{n_1}(X_i-\overline{X})^2,\quad S_2^2=\frac{1}{n_2-1}\sum_1^{n_2}(Y_i-\overline{Y})^2$$

易知

$$\overline{X}-\overline{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

于是在零假设 $H_0:\mu_1=\mu_2$ 下，随机变量

$$\xi\xlongequal{\mathrm{d}}\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

可见 $|\xi|$ 的值太大时应拒绝 H_0. 但 σ_1^2 和 σ_2^2 是未知的，ξ 不是统计量. 自然想到用 S_1^2 代替 σ_1^2，S_2^2 代替 σ_2^2. 于是应采用统计量

$$T=(\overline{X}-\overline{Y})\Big/\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\tag{4.25}$$

当 $|T|$ 太大时应拒绝 H_0. 应指出的是在 H_0 下 T 的精确分布相当复杂（而且依赖于比值 σ_1^2/σ_2^2）. 幸运的是，可以证明，在 H_0 下统计量 T 近似服从 m 个自由度的 t 分布. 这个 m 是与下列 m^* 最接近的整数.

$$m^*=\frac{\left(\frac{1}{n_1}S_1^2+\frac{1}{n_2}S_2^2\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}\tag{4.26}$$

利用 t 分布临界值表，找出临界值 λ 满足 $P(|T|>\lambda)=\alpha$，于是当且仅当 $|T|>\lambda$ 时拒绝 $H_0:\mu_1=\mu_2$.

类似地，也可解决 $\sigma_1^2\neq\sigma_2^2$ 时如何检验 $H_0:\mu_1\leqslant\mu_2$ 的问题.

例 4.6 研究患心脏病的父亲是否引起子女的胆固醇水平偏高的问题. 随机调查了 100 个 2 至 14 岁的孩子（其父皆死于心脏病），其胆固醇水平的平均值是 207.3，标准差是 35.6；另外，随机调查了父亲无心脏病史的 74 个 2 至 14 岁的孩子，其胆固醇水平的平均值是 193.4，标准差是 17.3，问：前者的胆固醇水平的平均值与后者的胆固醇水平的平均值有无显著差异？

设父亲死于心脏病的孩子的胆固醇水平 X 服从正态分布 $N(\mu_1,\sigma_1^2)$，父亲无心脏病史的孩子的胆固醇水平服从正态分布 $N(\mu_2,\sigma_2^2)$. 这里参数 $\mu_1,\sigma_1^2,\mu_2,\sigma_2^2$ 都是未知的. 我们要检验的假设是 $H_0:\mu_1=\mu_2$.

首先判别 σ_1^2 是否等于 σ_2^2. 使用统计量 $F=S_1^2/S_2^2$，这里 S_1^2 和 S_2^2 分别是两个样本的方差，经过计算知 $F=4.23$. 利用否定域(4.14). 对于 $\alpha=0.05$，查表得 $C_1=0.6548$，$C_2=1.5491$. 现在 $F>C_2$，故可以认为 $\sigma_1^2\neq\sigma_2^2$.

为了检验 $H_0:\mu_1=\mu_2$，使用统计量(4.25). 现在 $\overline{X}=207.3$，$\overline{Y}=193.4$. 代入后和统计量 T 的值为 3.40.

从(4.26)知 $m^*=151.4$. 于是，在 H_0 下统计量 T 近似服从 151 个自由度的 t 分布. 设 $t_{0.975}(l)$ 是 l 个自由度的 t 分布的 0.975 分位数，则在 H_0 下 $P(|T|>t_{0.975}(151))\doteq 0.05$. 故临界值 $\lambda=t_{0.975}(151)<t_{0.975}(120)=1.980$. 现在 $|T|=3.40>1.980$，因而应拒绝 $H_0:\mu_1=\mu_2$. 即可以认为胆固醇水平的平均值有显著性差异，父亲无心脏病史的孩子的胆固醇水平的平均值确实低些.

成对数据的比较

有些实际问题里的数据是天然成对的：$(x_1,y_1),\cdots,(x_n,y_n)$，其中 $x_1,\cdots,x_n$ 来自某总体，$y_1,\cdots,y_n$ 来自另一个总体，但 x_i 与 y_i 不能认为是相互独立的. 设 $Ex_i\equiv\mu_1\quad Ey_i\equiv\mu_2$（这里把 x_i,y_i 看成随机变量）. 要检验的假设是 $H_0:\mu_1=\mu_2$. 怎么办？令 $z_i=x_i-y_i(i=1,2,\cdots,n)$. 通常在实际工作中假定 $z_1,\cdots,z_n$ 是来自某个正态分布 $N(\mu_1-\mu_2,\sigma^2)$ 的简单随机样本（σ^2 未知）. 为了检验 $H_0:\mu_1=\mu_2$，可使用统计量

$$T=\frac{\overline{z}}{\sqrt{S^2/n}},\quad (参看(4.3))\tag{4.27}$$

这里 $\overline{z}=\dfrac{1}{n}\sum\limits_1^n z_i$，$S^2=\dfrac{1}{n-1}\sum\limits_1^n(z_i-\overline{z})^2$. 当 $|T|>c$ 时拒绝 H_0，

这里临界值 c 根据检验水平 α 来确定.

例 4.7 为了鉴定两种工艺方法对产品某性能指标有无显著性差异,对于 9 批材料用两种工艺进行生产,得到该指标的 9 对数据如下:

0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.10	0.21	0.52	0.32	0.78	0.59	0.68	0.77	0.89

问:根据上述数据,能否说两种不同工艺对产品的该性能指标有显著性差异?(检验水平 $\alpha=0.05$)

解 考查 9 对数据的差:$z_1,\cdots,z_9$. 使用统计量(4.27). 查 t 分布的临界值表得临界值 $\lambda=2.306$(自由度是 8). 现在统计量 T 的值是 $1.5<\lambda$. 故不能认为两种工艺下该指标的平均值有显著性差异.

以上我们对于一个正态总体、两个正态总体的假设检验问题进行了比较完全的讨论. 利用广义似然比检验法导出了各种检验问题的具体否定域(Behrens - Fisher 问题的否定域除外)这些否定域经过各自的数学研究,大都有某种优良性——一致最大功效无偏(UMPU). 为便于查阅,我们把正态总体情形下的各种检验问题、检验统计量及否定域列为表 4.1 与表 4.2.

在表 4.1 和表 4.2 中一总体情形,$X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的样本;在两总体情形,$X_1,\cdots,X_{n_1}$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本,$Y_1,\cdots,Y_{n_2}$ 是来自 $N(\mu_2,\sigma_2^2)$ 的样本. $\chi^2(x;n)$ 是 n 个自由度 χ^2 分布的密度函数,$t(x;n)$ 是 n 个自由度 t 分布的密度函数,$f(x;n_1,n_2)$ 是自由度为 n_1 和 n_2 的 F 分布的密度函数. UMP 表示否定域是一致最大功效的,UMPU 表示否定域是一致最大功效无偏的.

对于非正态总体,广义似然比检验法也是可用的. 要注意,不是在所有情况下广义似然比检验法都好.

例 4.8 设 X 取值限于集合 $\{2,-2,1,-1,0\}$,$P(X=x)=p(x,\theta)\ (0\leqslant\theta<1)$,其中 θ 是未知参数.

表 4.1 正态总体参

	零假设 H_0	备择假设 H_1	检验统计量	分　布
已知 $\sigma=\sigma_0$	$\mu\leqslant\mu_0$	$\mu>\mu_0$	$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	$N\left(\mu,\frac{\sigma_0^2}{n}\right)$
	$\mu=\mu_0$	$\mu\neq\mu_0$		
	$\mu\leqslant\mu_1$ 或 $\mu\geqslant\mu_2$	$\mu_1<\mu<\mu_2$		
	$\mu_1\leqslant\mu\leqslant\mu_2$	$\mu<\mu_1$ 或 $\mu>\mu_2$		
已知 $\mu=\mu_0$	$\sigma^2\leqslant\sigma_0^2$	$\sigma^2>\sigma_0^2$	$\chi^2=\sum_{i=1}^{n}(X_i-\mu_0)^2$	$\frac{1}{\sigma^2}\chi^2\sim\chi^2(n)$
	$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$		
	$\sigma^2\leqslant\sigma_1^2$ 或 $\sigma^2\geqslant\sigma_2^2$	$\sigma_1^2<\sigma^2<\sigma_2^2$		
	$\sigma_1^2\leqslant\sigma^2\leqslant\sigma_2^2$	$\sigma^2<\sigma_1^2$ 或 $\sigma^2>\sigma_2^2$		

* 其中 Φ 表示 $N(0,1)$ 分布的分布函数.

自由度	否定域	C 的决定	最优性
	$\overline{X} > C$	$\Phi\left(\frac{C-\mu_0}{\sigma_0}\sqrt{n}\right)=1-\alpha$	UMP
	$\overline{X} < C_1$ 或 $\overline{X} > C_2$	$\Phi\left(\frac{C_1-\mu_0}{\sigma_0}\sqrt{n}\right)=\frac{\alpha}{2}$ $C_2=2\mu_0-C_1$	UMPU
	$C_1 < \overline{X} < C_2$	$\Phi\left(\frac{C_2-\mu_1}{\sigma_0}\sqrt{n}\right)-\Phi\left(\frac{C_1-\mu_1}{\sigma_0}\sqrt{n}\right)$ $=\Phi\left(\frac{C_2-\mu_2}{\sigma_0}\sqrt{n}\right)-\Phi\left(\frac{C_1-\mu_2}{\sigma_0}\sqrt{n}\right)=\alpha$	UMP
	$\overline{X} < C_1$ 或 $\overline{X} > C_2$	$\Phi\left(\frac{C_2-\mu_1}{\sigma_0}\sqrt{n}\right)-\Phi\left(\frac{C_1-\mu_1}{\sigma_0}\sqrt{n}\right)$ $=\Phi\left(\frac{C_2-\mu_2}{\sigma_0}\sqrt{n}\right)-\Phi\left(\frac{C_1-\mu_2}{\sigma_0}\sqrt{n}\right)=1-\alpha$	UMPU
n	$\chi^2 > C\sigma_0^2$	$\int_C^{\infty}\chi^2(x;n)\mathrm{d}x=\alpha$	UMP
	$\chi^2 < C_1\sigma_0^2$ 或 $\chi^2 > C_2\sigma_0^2$	$\int_{C_1}^{C_2}\chi^2(x;n)\mathrm{d}x$ $=\frac{1}{n}\int_{C_1}^{C_2}\chi^2(x;n+2)\mathrm{d}x=1-\alpha$	UMPU
	$C_1 < \chi^2 < C_2$	$\int_{C_1/\sigma_1^2}^{C_2/\sigma_1^2}\chi^2(x;n)\mathrm{d}x$ $=\int_{C_1/\sigma_2^2}^{C_2/\sigma_2^2}\chi^2(x;n)\mathrm{d}x=\alpha$	UMP
	$\chi^2 < C_1$ 或 $\chi^2 > C_2$	$\int_{C_1/\sigma_1^2}^{C_2/\sigma_1^2}\chi^2(x;n)\mathrm{d}x$ $=\int_{C_1/\sigma_2^2}^{C_2/\sigma_2^2}\chi^2(x;n)\mathrm{d}x=1-\alpha$	UMPU

表 4.2 正态总体参数

<table>
<tr><th></th><th>零假设 H_0</th><th>备择假设 H_1</th><th>检验统计量</th><th>分布</th></tr>
<tr><td rowspan="6">一总体情形</td><td>$\sigma^2\leqslant\sigma_0^2$
$-\infty<\mu<\infty$</td><td>$\sigma^2>\sigma_0^2$
$-\infty<\mu<\infty$</td><td rowspan="4">$\chi^2=\sum_{i=1}^{n}(X_i-\overline{X})^2$</td><td rowspan="4">$\frac{1}{\sigma^2}\chi^2\sim\chi^2(n-1)$</td></tr>
<tr><td>$\sigma^2=\sigma_0^2$
$-\infty<\mu<\infty$</td><td>$\sigma^2\neq\sigma_0^2$
$-\infty<\mu<\infty$</td></tr>
<tr><td>$\sigma^2\leqslant\sigma_1^2$或$\sigma^2\geqslant\sigma_2^2$
$-\infty<\mu<\infty$</td><td>$\sigma_1^2<\sigma^2<\sigma_2^2$
$-\infty<\mu<\infty$</td></tr>
<tr><td>$\sigma_1^2\leqslant\sigma^2\leqslant\sigma_2^2$
$-\infty<\mu<\infty$</td><td>$\sigma^2<\sigma_1^2$或$\sigma^2>\sigma_2^2$
$-\infty<\mu<\infty$</td></tr>
<tr><td>$\mu\leqslant\mu_0$
$\sigma^2>0$</td><td>$\mu>\mu_0$
$\sigma^2>0$</td><td rowspan="2">$T=\dfrac{\sqrt{n}(\overline{X}-\mu_0)}{\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}}$</td><td rowspan="2">$t$</td></tr>
<tr><td>$\mu=\mu_0$
$\sigma^2>0$</td><td>$\mu\neq\mu_0$
$\sigma^2>0$</td></tr>
<tr><td rowspan="6">两总体情形</td><td>$\sigma_1^2/\sigma_2^2\leqslant\Delta_0$
$-\infty<\mu_1,$
$\mu_2<\infty$</td><td>$\sigma_1^2/\sigma_2^2>\Delta_0$
$-\infty<\mu_1,$
$\mu_2<\infty$</td><td rowspan="2">$F=\dfrac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}\cdot\dfrac{1}{\Delta_0}\cdot\dfrac{n_2-1}{n_1-1}$</td><td rowspan="2">$F$</td></tr>
<tr><td>$\sigma_1^2/\sigma_2^2=\Delta_0$
$-\infty<\mu_1,$
$\mu_2<\infty$</td><td>$\sigma_1^2/\sigma_2^2\neq\Delta_0$
$-\infty<\mu_1,$
$\mu_2<\infty$</td></tr>
<tr><td>$\mu_1\leqslant\mu_2$
$\sigma_1^2=\sigma_2^2>0$</td><td>$\mu_1>\mu_2$
$\sigma_1^2=\sigma_2^2>0$</td><td rowspan="2">$T=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}}\times\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}$</td><td rowspan="2">$t$</td></tr>
<tr><td>$\mu_1=\mu_2$
$\sigma_1^2=\sigma_2^2>0$</td><td>$\mu_1\neq\mu_2$
$\sigma_1^2=\sigma_2^2>0$</td></tr>
<tr><td>$\mu_1\leqslant\mu_2$
$\sigma_1^2\neq\sigma_2^2$</td><td>$\mu_1>\mu_2$
$\sigma_1^2\neq\sigma_2^2$</td><td rowspan="2">$T=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$
$S_1^2=\dfrac{1}{n_1-1}\sum_{1}^{n_1}(X_i-\overline{X})^2$
$S_2^2=\dfrac{1}{n_2-1}\sum_{1}^{n_2}(Y_i-\overline{Y})^2$</td><td rowspan="2">$t$(近似)</td></tr>
<tr><td>$\mu_1=\mu_2$
$\sigma_1^2\neq\sigma_2^2$</td><td>$\mu_1\neq\mu_2$
$\sigma_1^2\neq\sigma_2^2$</td></tr>
</table>

[注] $m^*=\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2\Big/\left[\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2\right]$.

的检验(二)

自由度	否定域	C 的 决 定	最优性
$n-1$	$\chi^2 > C\sigma_0^2$	$\int_C^{\infty} \chi^2(x;n-1)\mathrm{d}x = \alpha$	UMPU
	$\chi^2 < C_1\sigma_0^2$ 或 $\chi^2 > C_2\sigma_0^2$	$\int_{C_1}^{C_2} \chi^2(x;n-1)\mathrm{d}x = \frac{1}{n-1}\int_{C_1}^{C_2} x\chi^2(x;n-1)\mathrm{d}x = 1-\alpha$	UMPU
	$C_1 < \chi^2 < C_2$	$\int_{C_1/\sigma_1^2}^{C_2/\sigma_1^2} \chi^2(x;n-1)\mathrm{d}x = \int_{C_1/\sigma_2^2}^{C_2/\sigma_2^2} \chi^2(x;n-1)\mathrm{d}x = \alpha$	UMPU
	$\chi^2 < C_1$ 或 $\chi^2 > C_2$	$\int_{C_1/\sigma_1^2}^{C_2/\sigma_1^2} \chi^2(x;n-1)\mathrm{d}x = \int_{C_1/\sigma_2^2}^{C_2/\sigma_2^2} \chi^2(x;n-1)\mathrm{d}x = 1-\alpha$	UMPU
$n-1$	$T > C$	$\int_C^{\infty} t(x;n-1)\mathrm{d}x = \alpha$	UMPU
	$\|T\| > C$	$\int_C^{\infty} t(x;n-1)\mathrm{d}x = \frac{\alpha}{2}$	UMPU
(n_1-1,n_2-1)	$F > C$	$\int_C^{\infty} f(x;n_1-1,n_2-1)\mathrm{d}x = \alpha$	UMPU
	$F < C_1$ 或 $F > C_2$	$\int_{-\infty}^{C_1} f(x;n_1-1;n_2-1)\mathrm{d}x = \int_{C_2}^{\infty} f(x;n_1-1,n_2-1)\mathrm{d}x = \frac{\alpha}{2}$	UMPU (近似)
(n_1+n_2-2)	$T > C$	$\int_C^{\infty} t(x;n_1+n_2-2)\mathrm{d}x = \alpha$	UMPU
	$\|T\| > C$	$\int_C^{\infty} t(x;n_1+n_2-2)\mathrm{d}x = \frac{\alpha}{2}$	UMPU
m (与 m^* 最接近的整数)	$T > C$	$\int_C^{\infty} t(x;m)\mathrm{d}x = \alpha$	较优
	$\|T\| > C$	$\int_C^{\infty} t(x;m)\mathrm{d}x = \frac{\alpha}{2}$	较优

$$p(x,0)=\begin{cases}\dfrac{\alpha}{2}, & x=\pm 2\\ \dfrac{1-2\alpha}{2}, & x=\pm 1\\ \alpha, & x=0\end{cases}$$

$$p(x,\theta)=\begin{cases}\theta C, & x=-2\\ \dfrac{1-C}{1-\alpha}\left(\dfrac{1}{2}-\alpha\right), & x=\pm 1\\ \alpha\left(\dfrac{1-C}{1-\alpha}\right), & x=0\\ (1-\theta)C, & x=2\end{cases}\quad (0<\theta<1)$$

其中 $0<\alpha<\dfrac{1}{2},\dfrac{\alpha}{2-\alpha}<C<\alpha.$

研究检验问题：

$$H_0:\theta=0\leftrightarrow H_a:0<\theta<1$$

为了寻求广义似然比检验，先计算广义似然比．设样本是 X_1（样本量是 1），广义似然比为

$$\lambda(x)=\frac{\sup\limits_{0\leqslant\theta<1} p(x,\theta)}{p(x,0)}$$

易知

$$\lambda(2)=\frac{2C}{\alpha}\quad\left(\text{因为}\frac{\alpha}{2}<C\right)$$

$$\lambda(-2)=\frac{2C}{\alpha}$$

$$\lambda(1)=\lambda(-1)=\frac{1-C}{1-\alpha}$$

$$\lambda(0)=\frac{1-C}{1-\alpha}$$

水平是 α 的否定域为 $W_0=\{x:\lambda(x)>\lambda_0\}$，其中 λ_0 适合

$$P(X_1\in W_0|\theta=0)=\alpha$$

易知取 $\lambda_0=\dfrac{1-C}{1-\alpha}$时

$$P(\lambda(X_1) > \lambda_0 \mid H_0) = P(X_1 = \pm 2 \mid \theta = 0) = \alpha$$

故否定域是 $W_0 = \{2, -2\}$，功效函数

$$P\left(\lambda(X_1) > \frac{1-C}{1-\alpha} \middle| \theta\right) = P(X_1 = \pm 2 \mid \theta) = \theta C + (1-\theta)C$$

$$= C < \alpha \quad (0 < \theta < 1)$$

可见这个检验法不是无偏的. 下列检验法更好些，取否定域 $W_1 = \{0\}$，即当且仅当 $X_1 = 0$ 时拒绝 H_0.

此时 $P(X_1 \in W_1 \mid 0) = p(0,0) = \alpha$. 功效函数

$$P(X_1 \in W_1 \mid \theta) = p(0,\theta) = \alpha \frac{1-C}{1-\alpha} > \alpha \quad (0 < \theta < 1)$$

此检验法是无偏的.

作为本节的结尾，我们指出 F 分布是应用极广的分布. 除了在正态总体的检验问题中常用到 F 分布外，F 分布的分位数与 β 分布的分位数有密切关系.

引理 4.1 设 X 与 Y 相互独立且 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，则

$$U \stackrel{\mathrm{d}}{=} \frac{X}{X+Y}$$

的分布函数是 $\beta\left(x; \frac{m}{2}, \frac{n}{2}\right)$ （β 分布），这里

$$\beta(x; p, q) = \frac{1}{B(p,q)} \int_0^x u^{p-1}(1-u)^{q-1} \mathrm{d}u \quad (0 \leqslant x \leqslant 1)$$

$$p > 0, q > 0, B(p,q) = \int_0^1 u^{p-1}(1-u)^{q-1} \mathrm{d}u$$

证明：令 $V = X + Y$，直接计算 (U,V) 的联合密度，不难推知 U 的分布函数正好是上述的 $\beta\left(x; \frac{m}{2}, \frac{n}{2}\right)$. 详细叙述从略.

证毕.

引理 4.2 设 $\beta_\gamma(K_1, K_2)$ 是 β 分布 $\beta(x; K_1, K_2)$ 的 γ 分位数（K_1, K_2 是正整数），$F_\gamma(n_1, n_2)$ 是自由度为 n_1, n_2 的 F 分布的 γ 分位数，则

$$\beta_\gamma(K_1,K_2)=\left(1+\frac{K_2}{K_1}\cdot\frac{1}{F_\gamma(2K_1,2K_2)}\right)^{-1} \tag{4.28}$$

证明:在引理 4.1 中取 $m=2K_1, n=2K_2$ 且

$$F=\frac{K_2}{K_1}\cdot\frac{X}{Y}$$

则

$$U=\frac{F}{\frac{K_2}{K_1}+F}$$

且 F 服从自由度为 $2K_1, 2K_2$ 的 F 分布. 于是

$$\beta(x;K_1,K_2)=P(U\leqslant x)$$
$$=P\left(F\leqslant\frac{K_2}{K_1}\cdot\frac{x}{1-x}\right)\quad(0<x<1).$$

由此可见

$$F_\gamma(2K_1,2K_2)=\frac{K_2}{K_1}\cdot\frac{\beta_\gamma(K_1,K_2)}{1-\beta_\gamma(K_1,K_2)}$$

故(4.28)成立. 证毕.

公式(4.28)是很有用的. 设 X 服从伯努利分布, $P(X=1)=p=1-P(X=0)$, $0\leqslant p\leqslant 1$, $X_1,\cdots,X_n$ 是 X 的样本, 从第二章例 3.3 知 p 的 γ 水平置信下限是 $\underline{p}\left(\sum_{1}^{n}X_i\right)$, 这里 $\underline{p}(0)=0$, $\underline{p}(K)(K\geqslant 1)$ 满足方程

$$\frac{n!}{(K-1)!(n-K)!}\int_0^{\underline{p}(K)}x^{K-1}(1-x)^{n-K}\mathrm{d}x=1-\gamma$$

不难看出 $\underline{p}(K)=\beta_{1-\gamma}(K,n-K+1)$ (β 分布的分位数), 从(4.28)知

$$\underline{p}(K)=\left(1+\frac{n-K+1}{K}\cdot\frac{1}{F_{1-\gamma}(2K,2(n-K+1))}\right)^{-1}$$
$$=\left(1+\frac{n-K+1}{K}F_\gamma(2(n-K+1),2K)\right)^{-1} \tag{4.29}$$

所以可用 F 分布的分位数表来求 p 的置信下限.

*§5 临界值和 p 值

前面说过，给出一个检验法就是要给出一个否定域. 否定域 W 通常是由一个直观上有明确意义的统计量 $\varphi(X_1,\cdots,X_n)$ 来确定. 确定的方式可概括为两种.

第一种(单边情形)：

$$W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)>\lambda\} \tag{5.1}$$

第二种(双边情形)：

$$W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)<\lambda_1 \text{ 或 } \varphi(x_1,\cdots,x_n)>\lambda_2\} \tag{5.2}$$

这里 λ 叫作单边情形的临界值，λ_1 和 λ_2 叫作双边情形的临界值. 临界值是根据检验水平 α 来确定的. 通常，对于单边情形，应找 λ 满足

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)=\alpha \tag{5.3}$$

这里"$\theta\in\Theta_0$"是零假设 H_0，$P_\theta(A)$ 表示参数的真值是 θ 时事件 A 的概率. 对于这个 λ，(5.3)式表明否定域(5.1)的精确检验水平恰好是 α.

当满足(5.3)的 λ 不存在时，应选 λ 满足

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)<\alpha\leqslant\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\lambda) \tag{5.4}$$

此时，否定域(5.1)的检验水平不超过 α.

对于双边情形应选取 $\lambda_1<\lambda_2$ 满足

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)<\lambda_1)=\frac{\alpha}{2} \tag{5.5}$$

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda_2)=\frac{\alpha}{2} \tag{5.6}$$

对这样的 λ_1,λ_2，否定域(5.2)的检验水平不超过 α. 当这样的 λ_1 不存在时，应选 λ_1 满足

$$\begin{aligned}&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)<\lambda_1)<\frac{\alpha}{2}\\ \leqslant&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\leqslant\lambda_1)\end{aligned} \tag{5.7}$$

当这样的 λ_2 不存在时，应选取 λ_2 满足

$$\begin{aligned}&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda_2)<\frac{\alpha}{2}\\ \leqslant&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\lambda_2)\end{aligned} \tag{5.8}$$

此时否定域(5.2)的检验水平不超过 α.

上述根据检验水平 α 确定临界值 λ 或 λ_1 和 λ_2 从而获得否定域的方法，简称临界值方法. 这是本书用来确定否定域的基本方法. 读者容易看出，前面关于正态总体的假设检验的否定域都是用这个方法确定的.

我们指出，还有一种确定否定域的方法——p 值方法. 该方法可提供人们更多的信息.

先研究单边情形的否定域(5.1). 设 $x_1^0,\cdots,x_n^0$ 是样本值(已知的 n 个常数). 令

$$p(x_1^0,\cdots,x_n^0)=\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0)) \tag{5.9}$$

这里"$\theta\in\Theta_0$"就是零假设 H_0. $p(x_1^0,\cdots,x_n^0)$是 H_0 成立时统计量取值不小于 $\varphi(x_1^0,\cdots,x_n^0)$的最大概率.

定义 5.1 $p(x_1^0,\cdots,x_n^0)$(由(5.9)定义)叫作单边情形下样本值$(x_1^0,\cdots,x_n^0)$的 p 值.

p 值的重要性在于下列

引理 5.1 设对给定的 $\alpha\in(0,1)$，恰有一个 λ 满足

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)=\alpha \tag{5.10}$$

则 $\varphi(x_1^0,\cdots,x_n^0)>\lambda$ 的充要条件是 $p(x_1^0,\cdots,x_n^0)<\alpha$.

这个引理告诉我们，在条件(5.10)下，样本值$(x_1^0,\cdots,x_n^0)$落入检验水平为 α 的否定域(5.1)的充要条件是样本值的 p 值小于 α. 换句话说，当且仅当样本值的 p 值小于 α 时拒绝 H_0. 即否定域(5.1)可表示为$\{(x_1,\cdots,x_n):p(x_1,\cdots,x_n)<\alpha\}$. 这种用 p 值确定否定域的方法叫作 p 值方法. p 值方法的优点在于：不预先确定检验水平 α，从计算出的 p 值知道，对一切大于这个 p 值的 α，则拒绝 H_0 引起的错误其概率不超过 α.

p 值 $p(x_1^0,\cdots,x_n^0)$也可看作是样本$(x_1^0,\cdots,x_n^0)$与零假设 H_0 相容程度的度量. p 值越大，相容程度越高；反之，p 值越小，则相容程度越低. p 值小到一定程度则认为二者不相容了，即应拒绝 H_0. 当 p 值小于 α 时认为二者不相容，这时拒绝 H_0，引起的错误其概率不超过 α.

引理 5.1 的证明：设 $p(x_1^0,\cdots,x_n^0)<\alpha$，则 $\sup\limits_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0))<\alpha$. 从(5.10)知 $\varphi(x_1^0,\cdots,x_n^0)>\lambda$. 反过来，设 $\varphi(x_1^0,\cdots,x_n^0)>\lambda$，则有 $\varepsilon>0$ 使得 $\varphi(x_1^0,\cdots,x_n^0)-\varepsilon>\lambda$. 于是

$$p(x_1^0,\cdots,x_n^0)=\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0))$$
$$\leqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)>\varphi(x_1^0,\cdots,x_n^0)-\varepsilon)$$
$$<\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)=\alpha$$

证毕.

给定 $\alpha\in(0,1)$，不一定有 λ 满足(5.10)(当检验用的统计量 $\varphi(X_1,\cdots,X_n)$ 是离散型随机变量时常出现这种情况). 此时应考虑检验水平不超过 α 的检验. 我们可推广引理 5.1.

引理 5.2 设对给定的 $\alpha\in(0,1)$，有 λ 满足

$$\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)<\alpha$$
$$\leqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\lambda) \tag{5.11}$$

则 $\varphi(x_1^0,\cdots,x_n^0)>\lambda$ 的充要条件是

$$p(x_1^0,\cdots,x_n^0)<\alpha$$

证明：设 $\varphi(x_1^0,\cdots,x_n^0)>\lambda$，则

$$p(x_1^0,\cdots,x_n^0)=\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0))$$
$$\leqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)>\lambda)<\alpha$$

反之，若 $\varphi(x_1^0,\cdots,x_n^0)\leqslant\lambda$，则

$$p(x_1^0,\cdots,x_n^0)\geqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\lambda)\geqslant\alpha$$

证毕.

从引理 5.2 知，若 λ 满足(5.11)，则否定域 $W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)>\lambda\}$ 的检验水平小于 α，而且样本值 $(x_1^0,\cdots,x_n^0)$ 落入 W 的充要条件是该样本值的 p 值小于 α.

例 5.1 设 $X\sim N(\mu,\sigma^2)$，σ 未知，检验假设 $H_0:\mu\leqslant\mu_0$.(备择假设是 $H_1:\mu>\mu_0$). 此时应取否定域(参看(4.5))

$$W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)>\lambda_0\}$$

其中 $\varphi(x_1,\cdots,x_n)=(\bar{x}-\mu_0)/\sqrt{S^2/n}$，$\bar{x}=\frac{1}{n}\sum_1^n x_i$，$S^2=\frac{1}{n-1}\sum_1^n(x_i-\bar{x})^2$，$\lambda_0=t_{1-\alpha}(n-1)$($n-1$ 个自由度 t 分布的 $1-\alpha$ 分位数).

根据样本值 $x_1^0,\cdots,x_n^0$，我们可直接计算 p 值：

$$p(x_1^0,\cdots,x_n^0)=\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0))$$

这里 $\theta=(\mu,\sigma^2),\Theta_0=\{(\mu,\sigma^2):\mu\leqslant\mu_0,\sigma^2>0\}$.

易知

$$p(x_1^0,\cdots,x_n^0)=P_{\mu_0,\sigma^2}(T\geqslant\varphi(x_1^0,\cdots,x_n^0))$$

这里 T 是服从 $n-1$ 个自由度的 t 分布的随机变量. 因此 $p(x_1^0,\cdots,x_n^0)$ 可以算出. 从引理 5.1 知,当且仅当这个 p 值小于 α 时拒绝 H_0.

例如,为了检验假设 $H_0:\mu\leqslant 25$,根据样本值 $x_1^0,\cdots,x_{64}^0$ 可算出 $\overline{x}=25.9$, $S^2=17.3$. 故 $\varphi(x_1^0,\cdots,x_{64}^0)=1.731$. 可计算出 p 值

$$p(x_1^0,\cdots,x_{64}^0)=P(T\geqslant 1.731)=0.044<0.05.$$

故对于检验水平 $\alpha=0.05$ 应拒绝 H_0.

现在来研究双边情形的否定域(5.2). 我们不去具体确定 λ_1 和 λ_2,但从统计量 $\varphi(X_1,\cdots,X_n)$ 的直观意义和(5.5)、(5.6),可找到 λ_0 满足 $\lambda_1\leqslant\lambda_0<\lambda_2$. 设 $x_1^0,\cdots,x_n^0$ 是样本值,当 $\varphi(x_1^0,\cdots,x_n^0)\leqslant\lambda_0$ 时令

$$p(x_1^0,\cdots,x_n^0)=\min\{2\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\leqslant\varphi(x_1^0,\cdots,x_n^0)),1\} \tag{5.12}$$

当 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_0$ 时令

$$p(x_1^0,\cdots,x_n^0)=\min\{2\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0)),1\} \tag{5.13}$$

这里"$\theta\in\Theta_0$"是零假设 H_0.

定义 5.2 由(5.12)和(5.13)定义的 $p(x_1^0,\cdots,x_n^0)$ 叫作双边情形下样本值 $(x_1^0,\cdots,x_n^0)$ 的 p 值.

p 值的重要意义见下列引理.

引理 5.3 设对给定的 $\alpha\in(0,1)$,有唯一的 λ_1 和唯一的 λ_2 满足

$$\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)<\lambda_1)\leqslant\frac{\alpha}{2}\leqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\leqslant\lambda_1) \tag{5.14}$$

$$\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)>\lambda_2)\leqslant\frac{\alpha}{2}\leqslant\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\lambda_2) \tag{5.15}$$

则"$\varphi(x_1^0,\cdots,x_n^0)<\lambda_1$ 或 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_2$"成立的充要条件是 $p(x_1^0,\cdots,x_n^0)<\alpha$.

证明:设 $\varphi(x_1^0,\cdots,x_n^0)<\lambda_1$,则 $\varphi(x_1^0,\cdots,x_n^0)\leqslant\lambda_0$(这是由于 $\lambda_0\in[\lambda_1,\lambda_2)$).

于是从(5.12)知

$$
\begin{aligned}
p(x_1^0,\cdots,x_n^0) &\leqslant 2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n) \\
&\leqslant \varphi(x_1^0,\cdots,x_n^0)) \\
&\leqslant 2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)<\lambda_1-\varepsilon) \\
&<\alpha(\text{对适当小的 }\varepsilon>0).
\end{aligned}
$$

若 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_2$，则存在 $\varepsilon>0$ 使得 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_2+\varepsilon$. 于是

$$
\begin{aligned}
p(x_1^0,\cdots,x_n^0) &\leqslant 2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0)) \\
&\leqslant 2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)>\lambda_2+\varepsilon) \\
&<\alpha
\end{aligned}
$$

总之，只要 $\varphi(x_1^0,\cdots,x_n^0)$ 小于 λ_1 或大于 λ_2，则一定有 $p(x_1^0,\cdots,x_n^0)<\alpha$.

另一方面，设 $p(x_1^0,\cdots,x_n^0)<\alpha$. 若 $p(x_1^0,\cdots,x_n^0)\leqslant\lambda_0$，则

$$
2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\leqslant\varphi(x_1^0,\cdots,x_n^0))<\alpha.
$$

从(5.14)知 $\varphi(x_1^0,\cdots,x_n^0)<\lambda_1$；若 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_0$，则

$$
2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,\cdots,X_n)\geqslant\varphi(x_1^0,\cdots,x_n^0))<\alpha.
$$

从(5.15)知 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_2$. 因此，只要 $p(x_1^0,\cdots,x_n^0)<\alpha$，则一定有 $\varphi(x_1^0,\cdots,x_n^0)<\lambda_1$ 或 $\varphi(x_1^0,\cdots,x_n^0)>\lambda_2$，引理 5.3 证毕.

引理 5.3 告诉我们，为了检验假设 $H_0:\theta\in\Theta_0$，否定域 $W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)<\lambda_1$ 或 $\varphi(x_1,\cdots,x_n)>\lambda_2\}$ 的检验水平不超过 α（有时检验水平恰好是 α），而且在引理 5.3 的条件下，样本值 $(x_1^0,\cdots,x_n^0)$ 落入这个否定域的充要条件是样本值的 p 值小于 α. 这种用 p 值来确定否定域的方法仍叫做 p 值方法.

双边情形下 p 值方法的优点与单边情形下 p 值方法的优点是一样的. 兹不赘述.

p 值方法有很大优点，但也有麻烦之处：要根据样本值计算出相应的 p 值. 有时这种计算还相当复杂. 好在一些常见的假设检验问题里，p 值的计算程序已在流行的统计软件包（如 SAS）中给出. 使用这些软件容易算出 p 值.

例 5.2 设 $X\sim N(\mu,\sigma^2)$，μ,σ^2 均未知，为了检验假设 $H_0:\sigma^2=\sigma_0^2$（σ_0 已知），前面说过应使用统计量 $\varphi(X_1,\cdots,X_n)=\frac{1}{\sigma_0^2}\sum_1^n(X_i-\overline{X})^2$. 当 φ 值太小或

太大时应拒绝 H_0. 故应采用双边情形的否定域

$$W=\{(x_1,\cdots,x_n):\varphi(x_1,\cdots,x_n)<\lambda_1 \text{ 或 } \varphi(x_1,\cdots,x_n)>\lambda_2\}$$

如何计算样本值$(x_1^0,\cdots,x_n^0)$的 p 值呢?

从直观上看,如果 H_0 成立,则$\dfrac{\varphi(x_1^0,\cdots,x_n^0)}{n-1}$应该和 1 相差不太大. 故在 W 的表达式中λ_1 应小于 $n-1$,λ_2 应大于 $n-1$. 取$\lambda_0=n-1$. 则 $\lambda_1<\lambda_0<\lambda_2$. 注意,$\varphi(x_1^0,\cdots,x_n^0)\leqslant\lambda_0$ 的充要条件是 $S^2\leqslant\sigma_0^2$,这里 $S^2=\dfrac{1}{n-1}\sum\limits_1^n(x_i^0-\overline{x}_0)^2$ $\left(\overline{x}_0=\dfrac{1}{n}\sum\limits_1^n x_i^0\right)$.

于是,从定义 5.2 知 $S^2\leqslant\sigma_0^2$ 时

$$p(x_1^0,\cdots,x_n^0)=\min\{2\sup_{\theta\in\Theta_0}P_\theta(\varphi(X_1,\cdots,X_n)\leqslant\varphi_0),1\}$$

这里 $\theta=(\mu,\sigma^2)$,$\Theta_0=\{(\mu,\sigma^2):\mu \text{ 任意},\sigma^2=\sigma_0^2\}$,$\varphi_0=\varphi(x_1^0,\cdots,x_n^0)$

易知 $p(x_1^0,\cdots,x_n^0)=\min\{2P(\xi\leqslant\varphi_0),1\}$,这里随机变量 ξ 服从$n-1$个自由度的χ^2分布.

类似地,当 $S^2>\sigma_0^2$ 时

$$p(x_1^0,\cdots,x_n^0)=\min\{2P(\xi\geqslant\varphi_0),1\}$$

(ξ 及 φ_0 之含义同前).

根据例 4.2 中提供的 10 个数据,知

$$S^2=75.7>64=\sigma_0^2,$$

$$\varphi_0=\frac{1}{\sigma_0^2}\sum_1^{10}(x_i^0-\overline{x}_0)^2=10.65,$$

$P(\xi\geqslant\varphi_0)=0.30$(因为 ξ 服从 9 个自由度的χ^2 分布). 于是 p 值 $=0.60$. 可见对一切 $\alpha\leqslant 0.60$,在检验水平 α 下都不应拒绝假设 $H_0:\sigma^2=64$.

§6 比率的假设检验

设 X 服从两点分布(伯努利分布). X 取值 0 或 1 且 $P(X=1)=p=1-P(X=0)$,这里 p 是未知的. p 就是所谓的"比率". 当 1 表示成功,0 表示失败,p 就是成功率;当 1 表示有效,0 表示无效,p 就是有效率.

先讨论一个总体的问题. 此时，对 p 的假设检验问题主要有下列三个.

(1) $H_0: p \leqslant p_0 \leftrightarrow H_a: p > p_0$

(2) $H_0: p \geqslant p_0 \leftrightarrow H_a: p < p_0$

(3) $H_0: p = p_0 \leftrightarrow H_a: p \neq p_0$，

这里 H_0 是要检验的零假设(其中 p_0 已知)，H_a 是备择假设.

首先讨论上述问题(1). 设 X 的简单随机样本是 $X_1, \cdots, X_n$，如何检验 $H_0: p \leqslant p_0$?

很自然想到 p 的估计量 $\hat{p} = \frac{1}{n}\sum_1^n X_i$，当 n 较大时，$\hat{p}$ 应与 p 很接近. 当 $\hat{p}$ 比 p_0 大得多时应拒绝 H_0. 令 $S = \sum_1^n X_i$，故对固定的 n，当 S 足够大时应拒绝 H_0. 所以否定域是 $\left\{(x_1, \cdots, x_n): \sum_1^n x_i \geqslant c\right\}$，其中 c 是临界值. 设检验水平是 α，取 c 为满足下式的最小整数：

$$\sup_{p \leqslant p_0} P_p(S \geqslant c) \leqslant \alpha \tag{6.1}$$

这里 $P_p(A)$ 表示总体的参数是 p 时事件 A 的概率.

注意，S 服从二项分布，即

$$P_p(S = i) = C_n^i p^i (1-p)^{n-i} \quad (i = 0, 1, \cdots, n)$$

于是 $P_p(S \geqslant k) = \sum_{i=k}^n C_n^i p^i (1-p)^{n-i}$

$$= \frac{n!}{(k-1)!(n-k)!}\int_0^p u^{k-1}(1-u)^{n-k}\,\mathrm{d}u$$

($k \geqslant 1$，见第二章)

由于 $P_p(S \geqslant k)$ 是 p 的增函数，(6.1)化为

$$P_{p_0}(S \geqslant c) \leqslant \alpha \tag{6.2}$$

我们不去求这个 c(满足(6.2)的最小整数)而是另想办法判别事件“$S \geqslant c$”是否发生. 设样本值是 $x_1^0, \cdots, x_n^0$(n 个已知数). 令 $s_0 = \sum_1^n x_i^0$. 显然，从(6.2)知 $s_0 \geqslant c$ 的充要条件是

$$\sum_{i=s_0}^{n} C_n^i p_0^i (1-p_0)^{n-i} \leqslant \alpha \tag{6.3}$$

设方程

$$\sum_{i=s_0}^{n} C_n^i p^i (1-p)^{n-i} = \alpha \tag{6.4}$$

的根为 $p_\alpha(s_0)(s_0 \geqslant 1)$. 此外,规定 $p_\alpha(0)=0$. 易知 $p_0 \leqslant p_\alpha(s_0)$. 故当且仅当 $p_\alpha(s_0) \geqslant p_0$ 时拒绝 $H_0: p \leqslant p_0$.

怎样计算 $p_\alpha(s_0)$呢? 有下列公式(参看(4.29)):

$$p_\alpha(s_0) = \left(1 + \frac{n-s_0+1}{s_0} F_{1-\alpha}(2(n-s_0+1), 2s_0)\right)^{-1} \tag{6.5}$$

例 6.1 一种广泛使用的药治疗慢性支气管炎的有效率是 0.80. 现在一家制药公司推出一种新药,声称:治疗慢性支气管炎的有效率高于 0.80,但药价比广泛使用的那种药减少四分之一. 为了验证新药的有效率是否高于 0.80,收集了临床试验数据. 从使用新药的病人中随机抽查了 30 人,其中 27 人有效,3 人无效. 问:能否认为新药的有效率高于 0.80?

解 用 X 表示使用新药的效果. $X=1$ 表示有效,$X=0$ 表示无效. 现在样本量 $n=30$,样本值之和 $s_0=27$,设检验水平 $\alpha=0.05$. 从(6.5)知

$$\begin{aligned} p_\alpha(s_0) &= p_{0.05}(27) = \left(1 + \frac{4}{27} F_{0.95}(8, 54)\right)^{-1} \\ &= \left(1 + \frac{4}{27} \times 2.13\right)^{-1} = 0.76 < p_0 = 0.80 \end{aligned}$$

故不能拒绝 $H_0: p \leqslant p_0$,即没有理由说新药比旧药有更高的有效率.

顺便说一下,若 30 人中有 28 人有效,则可计算出 $p_{0.05}(28) = \left(1 + \frac{3}{28} F_{0.95}(6, 56)\right)^{-1} = 0.814 > p_0$. 故要拒绝 H_0,即可认为新药有更高的有效率.

现在来讨论假设 $H_0: p \geqslant p_0$ 的检验问题. 设 $X_1, \cdots, X_n$ 是简

单随机样本，很自然想到，当 $S=\sum_{1}^{n}X_i$ 太小时应拒绝 H_0. 对给定的 $\alpha\in(0,1)$. 设 c 是满足下式的最大整数.

$$\sup_{p\geqslant p_0} P_p(S\leqslant c)\leqslant\alpha \tag{6.6}$$

可以证明 $P_p(S\leqslant c)$ 是 p 的减函数，于是(6.6)化为

$$P_{p_0}(S\leqslant c)\leqslant\alpha \tag{6.7}$$

我们不去寻找 c 的具体数值，而是另想办法判别"$S\leqslant c$"是否发生. 仿效前一检验问题的处理方法，设样本值是 $x_1^0,\cdots,x_n^0$，令 $s_0=\sum_{1}^{n}x_i^0$. 不难看出 $s_0\leqslant c$ 的充要条件是 $P_{p_0}(S\leqslant s_0)\leqslant\alpha$，即

$$\sum_{i=0}^{s_0}\mathrm{C}_n^i p_0^i(1-p_0)^{n-i}\leqslant\alpha$$

这等价于

$$\sum_{i=s_0+1}^{n}\mathrm{C}_n^i p_0^i(1-p_0)^{n-i}\geqslant 1-\alpha$$

设方程

$$\sum_{i=s_0+1}^{n}\mathrm{C}_n^i p^i(1-p)^{n-i}=1-\alpha \tag{6.8}$$

的根为 $\tilde{p}_\alpha(s_0)$. 于是 $s_0\leqslant c$ 的充要条件是 $\tilde{p}_\alpha(s_0)\leqslant p_0$. 所以，当且仅当 $\tilde{p}_\alpha(s_0)\leqslant p_0$ 时应拒绝 $H_0:p\geqslant p_0$.

怎样计算 $\tilde{p}_\alpha(s_0)$ 呢? 从(6.8)看出，$\tilde{p}_\alpha(s_0)$ 是贝塔分布 $\beta(x;s_0+1,n-s_0)$ 的 $1-\alpha$ 分位数，于是

$$\tilde{p}_\alpha(s_0)=\left(1+\frac{n-s_0}{s_0+1}\frac{1}{F_{1-\alpha}(2(s_0+1),2(n-s_0))}\right)^{-1} \tag{6.9}$$

这里 $F_{1-\alpha}(n_1,n_2)$ 是自由度是 n_1,n_2 的 F 分布的 $1-\alpha$ 分位数(参看(4.28)).

现在来研究 $H_0:p_1=p_0$ 的检验问题. 设 $X_1,\cdots,X_n$ 是简单随机样本，$S=\sum_{1}^{n}X_i$. 显然，S 太大或太小应拒绝 H_0. 对给定的检验

水平 α,应取最大的整数 c_1 和最小的整数 c_2 满足:

$$P_{p_0}(S\leqslant c_1)\leqslant\frac{\alpha}{2},\quad P_{p_0}(S\geqslant c_2)\leqslant\frac{\alpha}{2}$$

设样本值是 $x_1^0,\cdots,x_n^0$. 令 $s_0=\sum_1^n x_i^0$. 不难看出,$s_0\leqslant c_1$ 的充要条件是 $P_{p_0}(S\leqslant s_0)\leqslant\frac{\alpha}{2}$,$s_0\geqslant c_2$ 的充要条件是 $P_{p_0}(S\geqslant s_0)\leqslant\frac{\alpha}{2}$. 于是 $s_0\leqslant c_1$ 的充要条件是 $\tilde{p}_{\frac{\alpha}{2}}(s_0)\leqslant p_0$(参看(6.9)),$s_0\geqslant c_2$ 的条件是 $p_{\frac{\alpha}{2}}(s_0)\geqslant p_0$(参看(6.5)),可见当且仅当 $\tilde{p}_{\frac{\alpha}{2}}(s_0)\leqslant p_0$ 或者 $p_{\frac{\alpha}{2}}(s_0)\geqslant p_0$ 时应拒绝 $H_0:p=p_0$. 这时检验水平不超过 α.

现在来研究两个总体的比较问题. 设 X 与 Y 相互独立,都服从伯努利分布,$P(X=1)=p_1=1-P(X=0)$,$P(Y=1)=p_2=1-P(Y=0)$,p_1,p_2 均未知,设 X 有简单随机样本 $X_1,\cdots,X_{n_1}$,Y 有简单随机样本 $Y_1,\cdots,Y_{n_2}$,考虑下列三个检验问题:

(4) $H_0:p_1\leqslant p_2\leftrightarrow H_a:p_1>p_2$

(5) $H_0:p_1\geqslant p_2\leftrightarrow H_a:p_1<p_2$

(6) $H_0:p_1=p_2\leftrightarrow H_a:p_1\neq p_2$

这里 H_0 是零假设,H_a 是备择假设.

令 $S_1=\sum_1^{n_1}X_i$,$S_2=\sum_1^{n_2}Y_i$. 则 p_1 和 p_2 的估计量分别是 $\hat{p}_1=\frac{S_1}{n_1}$,$\hat{p}_2=\frac{S_2}{n_2}$. 自然想到:当 $\hat{p}_1$ 比 $\hat{p}_2$ 大得多时应拒绝 $H_0:p_1\leqslant p_2$;当 $\hat{p}_1$ 比 $\hat{p}_2$ 小得多时应拒绝 $H_0:p_1\geqslant p_2$;当 $\hat{p}_1$ 与 $\hat{p}_2$ 相差较大时应拒绝 $H_0:p_1=p_2$.

基于上述思想,我们给出两种检验法. 一是正态理论方法,这是大样本情形才能用的近似方法,另一是 Fisher 精确检验法,各种情形下都可以用,但计算上比较复杂.

先介绍正态理论方法,易知 $\mathrm{Var}(\hat{p}_1-\hat{p}_2)=\frac{1}{n_1}p_1(1-p_1)+\frac{1}{n_2}p_2(1-p_2)$. 令

$$\xi=\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1)+\frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)}} \tag{6.10}$$

$$\eta=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1)+\frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)}} \tag{6.11}$$

$$\zeta=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}, \tag{6.12}$$

这里 $\hat{p}_1=\frac{1}{n_1}\sum_1^{n_1}X_i,\hat{p}_2=\frac{1}{n_2}\sum_1^{n_2}Y_i,$

$$\hat{p}=\frac{1}{n_1+n_2}(n_1\hat{p}_1+n_2\hat{p}_2)$$

数学上可以证明，当 n_1 和 n_2 相当大(一般要求 $n_1\hat{p}_1(1-\hat{p}_1)\geqslant 5$，$n_2\hat{p}_2(1-\hat{p}_2)\geqslant 5$)时，随机变量 ξ 近似服从标准正态分布. 给定检验水平 α，设 z_α 是标准正态分布的 α 分位数，则 $P(\xi>z_{1-\alpha})\doteq\alpha$. 易知在 $p_1\leqslant p_2$ 的假设下 $\xi\geqslant\eta$，更有 $P(\eta>z_{1-\alpha})\leqslant P(\xi>z_{1-\alpha})\doteq\alpha$. 于是 $\eta>z_{1-\alpha}$ 时拒绝 $H_0:p_1\leqslant p_2$；当 $\eta\leqslant z_{1-\alpha}$ 时不拒绝 H_0.

类似地，在 $p_1\geqslant p_2$ 时，$\xi\leqslant\eta$，从而 $P(\eta<z_\alpha)\leqslant P(\xi<z_\alpha)\doteq\alpha$. 于是当且仅当 $\eta<z_\alpha$ 时拒绝 $H_0:p_1\geqslant p_2$.

可以证明，在 $p_1=p_2$ 的条件下，只要 n_1,n_2 相当大(一般要求 $n_1\hat{p}_1(1-\hat{p}_1)\geqslant 5,n_2\hat{p}_2(1-\hat{p}_2)\geqslant 5$)，统计量 ζ(见(6.12))近似服从标准正态分布，从而 $P(|\zeta|>z_{1-\frac{\alpha}{2}})\doteq\alpha$. 当且仅当 $|\zeta|>z_{1-\frac{\alpha}{2}}$ 时拒绝 $H_0:p_1=p_2$.

例 6.2 研究口服避孕药对年龄在 40 至 44 岁的妇女心脏的影响. 收集的资料表明，在 5 000 个使用口服避孕药的妇女中三年内出现心肌梗死的有 13 人，而在 10 000 个不使用口服避孕药的妇女中三年内出现心肌梗死的有 7 人. 试问：口服避孕药是否对

妇女的心脏有显著的影响?

解 用 p_1 表示年龄在 40 至 44 岁的妇女由于口服避孕药导致三年内出现心肌梗死的概率,p_2 表示这个年龄段的妇女不服这种避孕药但在三年内出现心肌梗死的概率. 我们要检验的假设是 $H_0: p_1 = p_2$,备择假设是 $H_a: p_1 \neq p_2$.

使用统计量(6.12),现在

$$\hat{p}_1 = \frac{13}{5\,000}, \quad \hat{p}_2 = \frac{7}{10\,000}$$

$$\hat{p} = \frac{13+7}{15\,000}$$

由于 $n_1\hat{p}_1(1-\hat{p}_1) = 6.66 \geqslant 5$,$n_2\hat{p}_2(1-\hat{p}_2) = 6.995 \geqslant 5$. 故可用正态理论检验法. 可计算出 $\zeta = 3.01$. 设检验水平 $\alpha = 0.01$,查表知 $z_{1-\frac{1}{2}\alpha} = z_{0.995} = 2.58$. 既然 $\zeta = 3.01 > 2.58$,故应拒绝 $H_0: p_1 = p_2$,即可认为口服避孕药对 40 至 44 岁的妇女心脏有显著影响.

现在来介绍 Fisher 精确检验法. 此时对样本量无任何限制. 先介绍操作方法,然后介绍这个检验法是基于何种统计思想推导出来的. 仍设 $X_1, \cdots, X_{n_1}$ 是第一个总体的简单随机样本,$Y_1, \cdots, Y_{n_2}$ 是第二个总体的简单随机样本,p_1 和 p_2 分别是两个总体的参数,p_1 和 p_2 均未知. 令 $S_1 = \sum_1^{n_1} X_i$,$S_2 = \sum_1^{n_2} Y_i$. 设两个样本的样本值分别是 $x_1^0, \cdots, x_{n_1}^0$ 和 $y_1^0, \cdots, y_{n_2}^0$,令

$$s_1^0 = \sum_1^{n_1} x_i^0, \quad s_2^0 = \sum_1^{n_2} y_i^0, \quad t = s_1^0 + s_2^0, \tag{6.13}$$

为了检验 $H_0: p_1 \leqslant p_2$(备择假设是 $H_a: p_1 > p_2$),令

$$p_1(s_1^0) = \sum_{i \geqslant s_1^0} p(i), \tag{6.14}$$

这里

$$p(i)=\frac{\binom{n_1}{i}\binom{n_2}{t-i}}{\binom{n_1+n_2}{t}}\text{①}\quad (i=0,1,\cdots) \tag{6.15}$$

对给定的检验水平 α，当且仅当 $p_1(s_1^0)\leqslant\alpha$ 时拒绝 $H_0:p_1\leqslant p_2$.

为了检验 $H_0:p_1\geqslant p_2$（备择假设是 $H_a:p_1<p_2$），令

$$p_2(s_1^0)=\sum_{i\leqslant s_1^0}p(i) \tag{6.16}$$

（$p(i)$之定义见(6.15)）. 对给定的检验水平 α，当且仅当 $p_2(s_1^0)\leqslant\alpha$ 时拒绝 $H_0:p_1\geqslant p_2$.

为了检验 $H_0:p_1=p_2$（备择假设是 $H_a:p_1\neq p_2$），令

$$p_3(s_1^0)=2\min\Big(\sum_{i\leqslant s_1^0}p(i),\sum_{i\geqslant s_1^0}p(i)\Big) \tag{6.17}$$

则对给定的检验水平 α，当且仅当 $p_3(s_1^0)\leqslant\alpha$ 时拒绝 $H_0:p_1=p_2$.

上述检验法就是 Fisher 精确检验法，其中“精确”二字的含义是，这个检验法的检验水平保证不超过给定的 α. 复杂之处在于要计算各个 $p(i)$，在实际计算时要利用下列递推关系式：当 $p(i)>0$ 时有

$$p(i+1)=p(i)\frac{(n_1-i)(t-i)}{(i+1)(n_2-t+i+1)} \tag{6.18}$$

这个关系式根据 $p(i)$的定义甚易验证其成立. 后面还要介绍实际工作中采用的列联表变换法，用于计算所有的 $p(i)$.

现在问：上述 Fisher 精确检验法是基于什么统计思想而导出的呢？沿用前面的记号，数学上可以证明，如果假设 $H_0:p_1\leqslant p_2$ 成立，则在 $S_1+S_2=t$ 的条件下，$S_1\geqslant c$ 的条件概率的最大值是 $\sum_{i=c}^{n_1}p(i)$，即

---

① $\binom{m}{i}$就是组合数 C_m^i. 当 $i>m$ 或 $i<0$ 时规定 $\binom{m}{i}=0$.

$$\sup_{p_1 \leqslant p_2} P_{p_1 p_2}(S_1 \geqslant c \mid S_1 + S_2 = t) = \sum_{i=c}^{n_1} p(i) \qquad (6.19)$$

这里 $P_{p_1 p_2}(A \mid S_1 + S_2 = t)$表示两个总体的参数分别是 p_1, p_2 时在 $S_1 + S_2 = t$ 的条件下事件 A 的条件概率，$p(i)$的定义见(6.15).

(6.19)的数学证明较长，从略. 给定 $\alpha \in (0,1)$. 设 c 是满足 $\sum_{i=c}^{n_1} p(i) \leqslant \alpha$ 的最小整数. 则在 $S_1 + S_2 = t$ 的条件下 $S_1 \geqslant c$ 时应拒绝 $H_0: p_1 \leqslant p_2$. 注意，根据样本值 $x_1^0, \cdots, x_{n_1}^0, y_1^0, \cdots, y_{n_2}^0$，从(6.13)知事件"$S_1 + S_2 = t$"已发生，而"$S_1 \geqslant c$"当且仅当 $s_1^0 \geqslant c$ 时发生，显然 $s_1^0 \geqslant c$ 的充要条件是 $p_1(s_1^0) \leqslant \alpha$.（参看(6.13)）. 故当且仅当 $p_1(s_1^0) \leqslant \alpha$ 时拒绝 $H_0: p_1 \leqslant p_2$，这时检验水平不超过 α.

类似地，数学上可以证明

$$\sup_{p_1 \geqslant p_2} P_{p_1 p_2}(S_1 \leqslant c \mid S_1 + S_2 = t) = \sum_{i=0}^{c} p(i)$$

给定 $\alpha \in (0,1)$. 设 c 是满足 $\sum_{i=0}^{c} p(i) \leqslant \alpha$ 的最大整数. 从(6.16)知 $s_1^0 \leqslant c$ 的充要条件是 $p_2(s_1^0) \leqslant \alpha$. 故当且仅当 $p_2(s_1^0) \leqslant \alpha$ 时应拒绝 $H_0: p_1 \geqslant p_2$. 这时检验水平不超过 α.

数学上可以证明，

$$P_{p_1 p_2}(S_1 \leqslant c_1 \mid S_1 + S_2 = t) = \sum_{i=0}^{c_1} p(i)$$

$$P_{p_1 p_2}(S_1 \geqslant c_2 \mid S_1 + S_2 = t) = \sum_{i=c_2}^{n_1} p(i)$$

取最大的整数 c_1 满足 $\sum_{i=0}^{c_1} p(i) \leqslant \frac{\alpha}{2}$，再取最小的整数 c_2 满足 $\sum_{i=c_2}^{n_2} p(i) \leqslant \frac{\alpha}{2}$. 则在 $H_0: p_1 = p_2$ 成立且 $S_1 + S_2 = t$ 的条件下事件"$S_1 \leqslant c_1$ 或 $S_1 \geqslant c_2$"的条件概率不超过 α. 可见，在 $S_1 + S_2 = t$ 的条件

下 $S_1 \leqslant c_1$ 或 $S_1 \geqslant c_2$ 时应拒绝 $H_0: p_1 = p_2$. 根据样本值 $x_1^0, \cdots, x_{n_1}^0$ 及 $y_1^0, \cdots, y_{n_2}^0$ 和(6.13)知"$S_1 + S_2 = t$"已经发生,故 $s_1^0 \leqslant c_1$ 或 $s_1^0 \geqslant c_2$ 时应拒绝 $H_0: p_1 = p_2$. 显然 $s_1^0 \leqslant c_1$ 的充要条件是 $2\sum_{i=0}^{s_1^0} p(i) \leqslant \alpha$, $s_1^0 \geqslant c_2$ 的充要条件是 $2\sum_{i=s_1^0}^{n_1} p(i) \leqslant \alpha$. 故从(6.17)知 $p_3(s_1^0) \leqslant \alpha$ 时拒绝 $H_0: p_1 = p_2$. 这时检验水平不超过 α.

以上叙述了导出 Fisher 精确检验法的统计思想. 下面介绍实际工作中采用的用于计算所有 $p(i)$ 的具体方法. 先引进一个定义. 对给定的 n_1, n_2 和 t(见(6.12)),称非负整数组成的矩阵

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

是宜取的,若 $a+b=n_1, c+d=n_2, a+c=t$. 显然,宜取阵由其左上角元素 a 唯一确定. 左上角是 a 的阵称为 a 阵,用 $\mathbf{A}_a$ 表示,显然,若

$$\mathbf{A}_a = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad (b \geqslant 1, c \geqslant 1)$$

则

$$\mathbf{A}_{a+1} = \begin{pmatrix} a+1 & b-1 \\ c-1 & d+1 \end{pmatrix}$$

从(6.18)知

$$p(a+1) = p(a)\frac{bc}{(a+1)(d+1)} \tag{6.20}$$

(6.20)比(6.18)的好处在于:公式好记.

从 $p(i)$ 的定义(见(6.15))知,当 i 不满足下列不等式(6.21)时,$p(i)=0$.

$$n_0 \leqslant i \leqslant n^* \tag{6.21}$$

这里

$$n_0 = \max(0, n_2 - t), \quad n^* = \min(n_1, t) \tag{6.22}$$

先依次列出 $A_{n_0}, A_{n_0+1}, \cdots, A_{n^*}$，然后计算 $p(n_0)$，再利用(6.20)逐次计算 $p(n_0+1), \cdots$. 当然，对于检验 $H_0: p_1 \leqslant p_2$ 只需列出 $A_{s_1^0}, A_{s_1^0+1}, \cdots, A_{n^*}$，计算出相应的 $p(s_1^0), p(s_1^0+1), \cdots, p(n^*)$；对于检验 $H_0': p_1 \geqslant p_2$，只需列出 $A_{n_0}, A_{n_0+1}, \cdots, A s_1^0$，计算出相应的 $p(n_0), p(n_0+1), \cdots, p(s_1^0)$.

例 6.3 某公安局有两个刑侦组，在过去一年内第一组接手25件人命案，结果侦破了23件；第二组接手35件人命案，结果侦破了30件. 问：两个组的侦破能力有无差别？

解 设两个组的侦破率分别为 p_1 和 p_2，要检验的假设是 $H_0: p_1 = p_2$.（这里设 X, Y 都是二值随机变量，$X=1$ 表示第一组侦破成功，$X=0$ 表示未能侦破；$Y=1$ 表示第二组侦破成功，$Y=0$ 表示未能侦破. $p_1 = P(X=1), p_2 = P(Y=1)$.）我们采用 Fisher 精确检验法. 现在 $n_1 = 25, n_2 = 35, s_1^0 = 23, s_2^0 = 30, t = 53$（参看(6.13)）. 从(6.22)知 $n_0 = 0, n^* = 25$. 从(6.17)知

$$p_3(s_1^0) = 2\min\left(\sum_{i=0}^{23} p(i), \sum_{i=23}^{25} p(i)\right).$$

从(6.15)知 $p(23) = 0.252$，从(6.20)知

$$p(24) = p(23)\frac{2\times 30}{24\times 6} = 0.105$$

$$p(25) = p(24)\frac{1\times 29}{25\times 7} = 0.017$$

于是

$$\sum_{i=23}^{25} p(i) = 0.374,$$

$$\sum_{i=0}^{23} p(i) = 1 - \sum_{i=23}^{25} p(i) + p(23)$$
$$= 1 - 0.374 + 0.252 = 0.878.$$

从而

$$p_3(s_1^0) = 2\times 0.374 = 0.748 > 0.05.$$

于是在检验水平 $\alpha = 0.05$ 下不应拒绝 $H_0: p_1 = p_2$. 换句话说，没有

理由认为两个组在破案能力上有差别.

§7 拟合优度检验

前面的讨论总是假定总体分布属于某个确定的类型(例如正态分布),而对其未知参数或数字特征(例如期望、方差)进行假设检验. 怎么知道一个总体的概率分布属于某种类型呢? 有时还要问,怎么知道一个总体的分布函数是某个给定的函数$F_0(x)$呢?

这是十分重要的问题. 有时根据对事物本质的分析,利用概率论的知识,可以给予回答. 但是在很多情况下,只能从一大堆数据中去发现规律,判断总体的分布是什么样子,这就是所谓的拟合问题.

一般而言,总是先根据样本值(一批观测数据)用第二章中介绍的直方图法,推测出总体可能遵从的概率分布(或密度函数),然后再利用本节所讲的方法来检验该总体的分布函数是否真的就是给出的函数. 下面先介绍χ^2检验法,然后介绍柯氏检验法和其他检验法.

(一) χ^2检验法

先讲"简单假设"①的检验.

设 X 的分布函数$F(x)$是未知的. 给定分布函数 $F_0(x)$(即满足:不减、右连续且$\lim\limits_{x\to\infty} F_0(x)=1$, $\lim\limits_{x\to-\infty} F_0(x)=0$),我们要研究的检验问题是:

$$H_0:F(x)\equiv F_0(x)\leftrightarrow H_1:F(x)\not\equiv F_0(x)$$

χ^2检验法可叙述如下:设 $X_1,\cdots,X_n$ 是来自 X 的样本. 在实轴上取 m 个点:$t_1,t_2,\cdots,t_m$ $(t_1<t_2<\cdots<t_m)$,于是把实轴分成

---

① 若在假设 H_0 成立的条件下总体的分布函数是完全确定的,则称 H_0 是简单假设;反之,称 H_0 是复杂假设.

$m+1$段，第1段是$(-\infty, t_1]$，第2段是$(t_1, t_2]$，…，第$m+1$段是(t_m, ∞). 用ν_i表示$X_1, \cdots, X_n$中落入第i段的个数$(i=1,2,\cdots, m+1)$，$\frac{\nu_i}{n}$是频率. 用p_i表示X的值落入第i段的概率. 如果H_0成立，则p_i是可以算出来的，事实上，

$$p_1 = P(X \leqslant t_1) = F_0(t_1)$$

$$p_i = P(t_{i-1} < X \leqslant t_i) = F_0(t_i) - F_0(t_{i-1}) \quad (2 \leqslant i \leqslant m)$$

$$p_{m+1} = P(X > t_m) = 1 - F_0(t_m)$$

根据概率与频率的关系知，如果H_0成立而且n又比较大，那么$\frac{\nu_i}{n}$与p_i差不多，就是说$\left(\frac{\nu_i}{n} - p_i\right)$应该比较小，于是

$$V = \sum_{i=1}^{m+1} \left(\frac{\nu_i}{n} - p_i\right)^2 \frac{n}{p_i}$$

也应该比较小才合理. 这是因为$\frac{n}{p_i}$起了平衡作用. 对于很小的p_i而言，即使$\left(\frac{\nu_i}{n} - p_i\right)^2$很小，$\frac{\nu_i}{n}$与$p_i$也可能相对误差很大. 因此$\left(\frac{\nu_i}{n} - p_i\right)^2 \Big/ p_i$是用来刻画$\frac{\nu_i}{n}$与$p_i$接近程度的量.

我们就取V作为统计量. 由于样本$X_1, \cdots, X_n$可以看成随机变量，所以

$$V = \sum_{i=1}^{m+1} \frac{(\nu_i - np_i)^2}{np_i} \tag{7.1}$$

也是随机变量.

经过数学方面的研究，可以证明（本书从略）：在H_0成立的条件下有

$$\lim_{n \to \infty} P(V \leqslant x) = \int_{-\infty}^{x} g_m(y) \mathrm{d}y$$

这里$g_m(y)$是m个自由度的χ^2分布的密度函数. 换句话说，在H_0下，V近似服从χ^2_m分布. 给定$\alpha \in (0,1)$查χ^2分布表，可找到λ使

$$\int_{\lambda}^{\infty} g_m(y)\mathrm{d}y = \alpha$$

于是 $P(V>\lambda|H_0)=\alpha$，H_0 的否定域为 $W_0=\{V>\lambda\}$. 即根据样本值 $x_1,\cdots,x_n$ 计算出 V，若 $V>\lambda$ 就拒绝 $H_0:F(x)\equiv F_0(x)$；若 $V\leqslant\lambda$ 就接受 H_0.

上述方法就是著名的 χ^2 检验法，是 K. Pearson 于 1900 年提出的. 在实际使用时如何选取 m 及分点 $t_1,\cdots,t_m$ 呢？这并无硬性规定. 对于离散型分布，这些数的选取有比较自然的方法（见下面的例）. 对于连续型分布，有人建议和使用直方图法时的情形相似，取 $m\leqslant\frac{n}{5}$；a 为比所有 x_i 略小的数，b 为比所有 x_i 略大的数，将区间 $[a,b]$ $m+1$ 等分，可得分点 $t_1,t_2,\cdots,t_m$.

读者可以通过下面的例子掌握 χ^2 检验法.

例 7.1 某工厂近 5 年来发生了 63 次事故，按星期几分类如下

星　期	一	二	三	四	五	六
次　数	9	10	11	8	13	12

问：事故的发生是否与星期几有关？

用 X 表示这样的随机变量：若事故发生在星期 i，则 $X=i$，显然 X 的可能值是 $1,2,\cdots,6$（星期日为该厂厂休日）.

我们要检验的假设是 $H_0:P(X=i)=\frac{1}{6}(i=1,2,\cdots,6)$.

使用上述统计量 $V=\sum_{i=1}^{6}\left(\nu_i-\frac{n}{6}\right)^2\Big/\frac{n}{6}$，这里 ν_i 是发生在星期 i 的事故次数（注意 $m=5$）. 在 H_0 下 V 近似服从 5 个自由度的 χ^2 分布. 查 χ^2 分布表知 $P(V>11.07)=0.05$ 现在 $\nu_1=9$，$\nu_2=10$，$\cdots$，$\nu_6=12$，计算得 $V=1.67<11.07$，故假设 H_0 不能拒绝，即不能认为发生事故与星期几有关.

通常在拟合问题中，要考虑的分布常含有未知参数. 也就是

说，待检验的假设是

$$H_0: F(x) \in \{F_0(x;\theta_1,\cdots,\theta_k):(\theta_1,\cdots,\theta_k)\in\Theta_0\}$$

这里 Θ_0 是已知的集合. 例如要检验 X 是否服从正态分布，也即要检验 $H_0: F(x)\in\left\{\Phi\left(\frac{x-\mu}{\sigma}\right),\mu\text{ 任意},\sigma>0\right\}$，其中 Φ 是标准正态分布函数.

这是复杂假设的问题. 自然想到也用前面介绍的 χ^2 检验法. 要注意的是，在 H_0 下 X 取值落入第 i 段的概率 p_i 无法直接算出. 此时可用下面的方法

和前面一样将 $(-\infty,\infty)$ 分为 $m+1$ 个区间 $(-\infty,t_1]$, $(t_1,t_2]$, $\cdots$, (t_m,∞)，则有

$$p_1(\theta_1,\cdots,\theta_k)=F_0(t_1;\theta_1,\cdots,\theta_k)$$

$$p_i(\theta_1,\cdots,\theta_k)=F_0(t_i;\theta_1,\cdots,\theta_k)-F_0(t_{i-1};\theta_1,\cdots,\theta_k)$$

$$i=2,\cdots,m$$

$$p_{m+1}(\theta_1,\cdots,\theta_k)=1-F_0(t_m;\theta_1,\cdots,\theta_k)$$

由样本 $X_1,\cdots,X_n$ 落入 $m+1$ 个区间的频率 $\frac{\nu_i}{n}$，$i=1,\cdots,m+1$，按多项分布可求得 $\theta_1,\cdots,\theta_k$ 的最大似然估计 $\theta_{10},\cdots,\theta_{k0}$，即由

$$\sum_{i=1}^{m+1}\frac{\nu_i}{p_i(\theta_1,\cdots,\theta_k)}\frac{\partial p_i(\theta_1,\cdots,\theta_k)}{\partial\theta_j}=0,\quad j=1,\cdots,k$$

求出 $\theta_{10},\cdots,\theta_{k0}$，记

$$\hat{p}_i=p_i(\theta_{10},\cdots,\theta_{k0})\quad(i=1,\cdots,m+1)$$

数学上可以证明：

$$V=\sum_{i=1}^{m+1}\frac{(\nu_i-n\hat{p}_i)^2}{n\hat{p}_i}$$

近似地服从 $m-k$ 个自由度的 χ^2 分布. 这里 k 是未知参数的个数. 但实际应用时，当 $F_0(\cdot)$ 是连续函数时，上述求 $\theta_{10},\cdots,\theta_{k0}$ 较麻烦，而变通地采用下述做法：假设总体的分布函数是 $F_0(x;\theta_1,\cdots,\theta_k)$，样本是 $X_1,\cdots,X_n$，首先用最大似然估计法找出未知参数

$\theta_1,\cdots,\theta_k$的估计值$\hat{\theta}_1,\cdots,\hat{\theta}_k$，于是得分布函数$F_0(x;\hat{\theta}_1,\cdots,\hat{\theta}_k)$. 形式上和前面一样，计算$p_i$：

$$p_1 = F_0(t_1;\hat{\theta}_1,\cdots,\hat{\theta}_k)$$

$$p_i = F_0(t_i;\hat{\theta}_1,\cdots,\hat{\theta}_k) - F_0(t_{i-1};\hat{\theta}_1,\cdots,\hat{\theta}_k), \quad i=2,3,\cdots,m$$

$$p_{m+1} = 1 - F_0(t_m;\hat{\theta}_1,\cdots,\hat{\theta}_k)$$

仍用统计量

$$V = \sum_{i=1}^{m+1} \frac{(\nu_i - np_i)^2}{np_i}$$

查χ^2_{m-k}分布表可找到λ满足

$$\int_\lambda^\infty g_{m-k}(y)\mathrm{d}y = \alpha$$

取否定域$W_0=\{V>\lambda\}$，即当$V>\lambda$时拒绝H_0，当$V\leqslant\lambda$时接受H_0. 这样的检验的第一类错误概率$P\{V>\lambda\mid H_0\}$也近似等于α.

例 7.2 历史上有人对一块放射性物质进行观测，记录每20秒内放射的α粒子数，共作了500次观测. 得表7.1：

表 7.1

放射出的粒子数	观测到的次数
0	14
1	35
2	70
3	105
4	102
5	81
6	52
7	23
8	8
9	5
不小于 10	5
合　　计	500

由文献知放射性物质一定时间间隔放射的粒子数一般遵从Poisson分布. 问:从这批数据出发能否认为放射的粒子数服从Poisson 分布?

设该物质在20秒的间隔内放射出的α粒子数为X,X的分布是$p(x)=P(X=x)(x=0,1,2,\cdots)$. 记

$$p_0(x,\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}\quad(x=0,1,2,\cdots)$$

要检验的假设是

$$H_0:p(x)\in\{p_0(x,\lambda):\lambda>0\}$$

首先,利用数据可找出λ的最大似然估计为$\hat{\lambda}=3.87$. 利用公式

$$p_i=\frac{\hat{\lambda}^i}{i!}e^{-\hat{\lambda}}\quad(i=0,1,2,\cdots,9)$$

及

$$p_{10}=\sum_{l=10}^{\infty}e^{-\hat{\lambda}}\frac{\hat{\lambda}^l}{l!}=1-\sum_{i=0}^{9}p_i$$

可算出$p_0,\cdots,p_{10}$,再计算统计量

$$V=\sum_{i=0}^{10}\frac{(\nu_i-np_i)^2}{np_i}$$

这里ν_i是n个观测值中α粒子数恰为i个的个数$(0\leqslant i\leqslant 9)$,$\nu_{10}$是粒子数不小于10的观测次数. 可算得$V=7.275$. 查$\chi^2$分布表(自由度是10−1=9)得到临界值$\lambda=16.9(\alpha=0.05)$. 现在$V=7.275<16.9$,故不能拒绝$H_0$. 即可以认为$\alpha$粒子数服从Poisson分布.

例 7.3 某车间生产滚珠,随机抽取了50个产品,测得它们的直径为(单位:mm):

15.0	15.8	15.2	15.1	15.9	14.7	14.8	15.5
15.6	15.3	15.1	15.3	15.0	15.6	15.7	15.8
14.5	14.2	14.9	14.9	15.2	15.0	15.3	15.6
15.1	14.9	14.2	14.6	15.8	15.2	15.9	15.2
15.0	14.9	14.8	14.5	15.1	15.5	15.5	15.1

15.1　15.0　15.3　14.7　14.5　15.5　15.0　14.7

14.6　14.2

我们问:滚珠直径是否服从正态分布?

设滚珠直径为 X,其分布是 $F(x)$,待检验的假设是 H_0:

$$F(x)\in\left\{\Phi\left(\frac{x-\mu}{\sigma}\right):\mu \text{ 任意},\sigma>0\right\}$$

仍用 χ^2 检验法. 首先从数据出发找出 μ,σ^2 的最大似然估计是

$$\hat{\mu}=15.1,\quad \sigma^2=(0.4325)^2$$

取 $m=6$,这 50 个数据最小的是 14.2,最大的是 15.9. 取 $a=14.05$,$b=16.15$,将 $[a,b]$ 7 等分,得分点 $t_1=14.35$,$t_2=14.65$,$t_3=14.95$,$t_4=15.25$,$t_5=15.55$,$t_6=15.85$,实数轴被这些 t_i 分成了 7 段,利用 $\Phi\left(\frac{x-15.1}{0.4325}\right)$ 及正态分布表可以得到 p_i 如下:

$$p_1=0.0414,\quad p_2=0.1077,\quad p_3=0.2154$$

$$p_4=0.2710,\quad p_5=0.2154,\quad p_6=0.1077$$

$$p_7=0.0414$$

现在来计算统计量 V,为便于检查,列表(表 7.2)如下:

表 7.2

i	1	2	3	4	5	6	7
p_i	0.041 4	0.107 7	0.215 4	0.271 0	0.215 4	0.107 7	0.041 4
np_i	2.070	5.385	10.770	13.550	10.770	5.385	2.070
v_i	3	5	10	16	8	6	2
(np_i-v_i)	0.864 9	0.148 2	0.592 5	6.002 5	7.672 9	0.378 2	0.004 9
$\frac{(np_i-v_i)^2}{np_i}$	0.417 8	0.027 5	0.055 1	0.443 0	0.712 4	0.070 2	0.002 4

$$V=\sum_{i=1}^{7}\frac{(np_i-v_i)^2}{np_i}=1.7284$$

取 $\alpha=0.05$，查χ^2分布表（自由度 $6-2=4$）得 $\lambda=9.49$. 现在 $V=1.7284<9.49$，故不能拒绝 H_0.

上面介绍的检验法称为分布函数的χ^2检验法，也称为拟合优度χ^2检验法. 它的好处在于不管总体的分布是什么都可以用，因而应用较广.

(二) 列联表的独立性检验

作为χ^2检验法的一种应用，现在介绍列联表的独立性检验.

设随机变量 X 的可能值是 $1,2,\cdots,s$；Y 的可能值是 $1,2,\cdots,t$ $(s\geqslant 2,t\geqslant 2)$. 现在对向量 (X,Y) 进行了 n 次独立观测，发现"X 取 i，Y 取 j"的次数是 n_{ij} $(i=1,2,\cdots,s,j=1,2,\cdots,t)$. 要据此检验下列假设：

$$H_0:X\text{ 与 }Y\text{ 相互独立}$$

在这种问题中常将数据排列成下表（表 7.3），称为 $s\times t$ 联立表或列联表.

表 7.3

X \ Y	1	2	$\cdots$	t	
1	n_{11}	n_{12}	$\cdots$	n_{1t}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2t}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
s	n_{s1}	n_{s2}	$\cdots$	n_{st}	$n_{s\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot t}$	n

记

$$n_{i\cdot}=\sum_{j=1}^{t}n_{ij}\quad(i=1,\cdots,s)$$

$$n_{\cdot j}=\sum_{i=1}^{s}n_{ij}\quad (j=1,\cdots,t)$$

当然

$$\sum_{j}n_{\cdot j}=\sum_{i}n_{i\cdot}=n$$

记

$$p_{ij}=P(X=i,Y=j),\quad p_i=P(X=i),\quad q_j=P(Y=j)$$

待检验的假设就是

$$H_0: p_{ij}=p_iq_j\quad (一切\ i,j) \tag{7.2}$$

首先在 H_0 成立的条件下找 p_i,q_j 的最大似然估计.

注意似然函数为 $L=\prod_{i=1}^{s}\prod_{j=1}^{t}(p_iq_j)^{n_{ij}}$，于是

$$\ln L=\sum_{i=1}^{s}\sum_{j=1}^{t}(n_{ij}\ln p_i+n_{ij}\ln q_j)$$

解方程组

$$\frac{\partial \ln L}{\partial p_i}=0,\frac{\partial \ln L}{\partial q_j}=0\quad (i=1,\cdots,s-1,j=1,\cdots,t-1)$$

$$p_s=1-\sum_{1}^{s-1}p_i,q_t=1-\sum_{1}^{t-1}q_j$$

不难知道最大似然估计为

$$\hat{p}_i=\frac{n_{i\cdot}}{n},\hat{q}_j=\frac{n_{\cdot j}}{n}\quad (i=1,\cdots,s,j=1,\cdots,t)$$

研究统计量

$$V=\sum_{i,j}\frac{(n_{ij}-n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} \tag{7.3}$$

根据(一)中介绍的一般理论，在 H_0 下，V 的极限分布是 l 个自由度的χ^2分布，其中 l=组数$-1-$未知参数个数$=st-1-(s+t-2)=(s-1)(t-1)$. 于是 V 近似服从$\chi^2((s-1)(t-1))$分布. 给定 $\alpha\in(0,1)$，查χ^2分布表，可找到临界值 λ 使 $P(V>\lambda)\doteq\alpha$. 取否定域 $W_0=\{V>\lambda\}$就可以对假设 H_0 进行检验了.

引理 7.1 V 有下列简单的计算公式：

$$V=n\left(\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right) \tag{7.4}$$

证明：

$$V=\sum_i\sum_j\left(n_{ij}-n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}\right)^2\Big/n\cdot\frac{n_{i\cdot}}{n}\cdot\frac{n_{\cdot j}}{n}$$

$$=n\sum_i\sum_j\left(n_{ij}-\frac{1}{n}n_{i\cdot}n_{\cdot j}\right)^2\Big/n_{i\cdot}n_{\cdot j}$$

$$=n\sum_i\sum_j\left(n_{ij}^2-\frac{2}{n}n_{ij}n_{i\cdot}n_{\cdot j}+\frac{1}{n^2}n_{i\cdot}^2n_{\cdot j}^2\right)\Big/n_{i\cdot}n_{\cdot j}$$

$$=n\left[\sum_i\sum_j\left(\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-\frac{2}{n}n_{ij}+\frac{1}{n^2}n_{i\cdot}n_{\cdot j}\right)\right]$$

$$=n\left[\sum_i\sum_j\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right].$$

证毕.

$s=t=2$ 时用得最广，叫 2×2 联立表（四格表）. 此时

$$V=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot1}n_{\cdot2}} \tag{7.5}$$

$$\text{自由度}=(s-1)(t-1)=1$$

例 7.4 为了探讨吸烟与患慢性支气管炎有无关系，调查了339人. 情况见下表（表 7.4）：

表 7.4

人　数	患慢性支气管炎	未患慢性支气管炎	合计
吸　烟	43	162	205
不吸烟	13	121	134
合　计	56	283	339

我们可以设想有两个随机变量 X,Y. 任何一个人，若他吸烟，令 $X=1$，若他不吸烟，则令 $X=2$，若他患慢性支气管炎，令 $Y=1$，若未患慢性支气管炎，则令 $Y=2$. 记

$$p_{ij}=P(X=i,Y=j)$$

$$p_i=P(X=i),q_j=P(Y=j)\quad(i,j=1,2)$$

我们来检验 H_0：

$$p_{ij}=p_iq_j\quad(\text{一切 } i,j)$$

根据上面介绍的χ^2检验法，首先计算统计量 V，利用公式(7.5)知

$$V=339\times\frac{(43\times121-13\times162)^2}{56\times283\times205\times134}=7.48$$

对 $\alpha=0.01$，查χ^2分布表(自由度为 1)知临界值 $\lambda=6.63$. 现在 $V=7.48>6.63$，故应拒绝 H_0，也即患慢性支气管炎与吸烟有关.

(三) 柯尔莫戈罗夫(Kolmogorov A N.)检验法

前面说过，不管总体的分布是什么类型，χ^2检验法都可以用. 不过，对于连续型的随机变量，本段介绍的柯氏检验法更好些.

设 X 的分布函数 $F(x)$是未知的，$X_1,\cdots,X_n$ 是样本. $F_0(x)$ 是给定的某个分布函数. 我们来研究下列检验问题：

$$H_0:F(x)\equiv F_0(x)\leftrightarrow H_a:F(x)\not\equiv F_0(x)$$

首先从样本出发求出经验分布函数 $F_n(x)$如下：

$$F_n(x)=\begin{cases}0, & x<X_{(1)}\\ \dfrac{k}{n}, & X_{(k)}\leqslant x<X_{(k+1)}\quad(k=1,2,\cdots,n-1)\\ 1, & x\geqslant X_{(n)}\end{cases}$$

这里 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 是样本 $X_1,\cdots,X_n$ 从小到大的重排. 计算分歧度

$$D_n=\sup_{-\infty<x<\infty}|F_n(x)-F_0(x)|$$

把样本看成随机变量，则 D_n 是随机变量. 从第二章定理 4.1 (Glivenko - Cantelli 定理)知，如果 H_0 成立，则 $P(\lim\limits_n D_n=0)=1$. 换句话说，如果 H_0 成立，而 n 又比较大，则 D_n 的值不应该太大. 如果 D_n 的值太大，则应该否定 H_0. 因而否定域应取

$$W_0=\{D_n>\lambda\}\tag{7.6}$$

给定 $\alpha\in(0,1)$后，应如何取 λ 使样本落入 W_0 的概率为 α? 这就涉

及到如何求出 D_n 的分布. 下列引理告诉我们,D_n 的计算公式相当简单,而且当 $F_0(x)$是连续函数时 D_n 的分布与$F_0(x)$无关.

引理 7.2

$$D_n=\max_{1\leqslant k\leqslant n}\max\left(\frac{k}{n}-F_0(x_{(k)}),F_0(x_{(k)}-0)-\frac{k-1}{n}\right) \tag{7.7}$$

而且如果 $F_0(x)$是连续函数,$x_1,\cdots,x_n$ 是来自 $F_0(x)$的样本,则 D_n 的分布与 $F_0(x)$无关.

证明:约定 $x_{(0)}\overset{\mathrm{d}}{=\!=}-\infty,x_{(n+1)}\overset{\mathrm{d}}{=\!=}\infty,F_0(-\infty)=F_n(-\infty)\overset{\mathrm{d}}{=\!=}0$. 则

$$\begin{aligned}D_n&=\max_{0\leqslant k\leqslant n}\sup_{x\in[x_{(k)},x_{(k+1)})}|F_n(x)-F_0(x)|\\&=\max_{1\leqslant k\leqslant n}\max\left(\frac{k}{n}-F_0(x_{(k)}),F_0(x_{(k)}-0)-\frac{k-1}{n}\right)\end{aligned}$$

这表明(7.7)成立. 若 $F_0(x)$连续,则 $F_0(x_1),F_0(x_2),\cdots,F_0(x_n)$相互独立,都服从[0,1]上的均匀分布,而 $F_0(x_{(1)})\leqslant F_0(x_{(2)})\leqslant\cdots\leqslant F_0(x_{(n)})$是均匀分布的次序统计量,它们的联合分布完全被均匀分布确定,与 $F_0(x)$无关,从而 D_n 的分布与$F_0(x)$无关.

证毕.

D_n 的分布有复杂的表达式,当 n 不太大时,D_n 的分位数表已造好,见本书附表 7;当 n 很大时,需要利用 D_n 的渐近分布. 柯氏在 1933 年证明了下列深刻的极限定理:

定理 7.1 设 X 的分布函数是连续函数 $F_0(x)$. 则

$$\lim_{n\to\infty}P(\sqrt{n}D_n\leqslant x)=Q(x)\quad(\text{一切 }x)$$

这里

$$Q(x)=\begin{cases}\sum_{k=-\infty}^{\infty}(-1)^k\mathrm{e}^{-2k^2x^2}, & x>0\\0, & x\leqslant 0\end{cases} \tag{7.8}$$

这个定理的证明相当长,超出了本课程的范围,读者可参看文献[2].

从这个定理知道，只要 H_0 成立，n 又比较大，则$\sqrt{n}D_n$ 的分布近似地就是 $Q(x)$. $Q(x)$的数值已造好了表，给定 $\alpha\in(0,1)$，可查表找到 λ 使

$$P(D_n > \lambda) \doteq P(\sqrt{n}D_n > \sqrt{n}\lambda)$$
$$\approx 1 - Q(\sqrt{n}\lambda) = \alpha$$

(在附表 7 中对 $\alpha=0.05,0.01$ 给出了 λ 的值). 否定域是

$$W_0 = \{D_n > \lambda\}$$

上述方法就是著名的柯氏检验法.

例 7.5 一批小动物体重为(g):

356.4	362.5	394.7	356.0	387.6
305.1	385.1	383.2	346.6	314.2
394.8	370.7	370.8	434.2	365.2
377.1	365.9	384.4	297.4	404.3
412.0	349.1	344.5		

试用柯氏检验法检验这批资料是否符合 $N(370.6,(29.1)^2)$($\alpha=0.05$).

设 $F_0(x)$为 $N(370.6,(29.1)^2)$的分布函数. 利用公式(7.7)可以算出 $D_n=0.104$，对 $\alpha=0.05$，$n=23$，可找到临界值 $\lambda=0.275$，现在 $D_n=0.104<0.275$ 故可以认为这批资料符合 $N(270.6,(29.1)^2)$.

要注意的是，若 $F_0(x)$的表达式中含有未知参数，则柯氏检验法不能直接用. 实际工作者往往从样本出发找出未知参数的最大似然估计值，代入 $F_0(x)$的表达式，从而得到确定的分布函数，再计算分歧度 D_n，仍采用上述的柯氏检验法进行检验. 这是一种近似处理法. 此时，$\sqrt{n}D_n$ 的极限分布是否仍是上述的 $Q(x)$就不得而知了.

关于分布拟合的检验，还有许多别的检验法. 比较著名的有 Cramer - Von Mises 检验法(1936). 它要检验的假设是 $H_0:F(x)\equiv F_0(x)$. 取统计量

$$W^2 = n\int_{-\infty}^{\infty}[F_n(x) - F_0(x)]^2 \mathrm{d}F_0(x)$$

这里 $F_n(x)$是 $X_1,\cdots,X_n$ 的经验分布函数，积分是 Stieltjes 积分，乃是一种加权平均. 显然，如果 H_0 成立，n 又比较大，则 W^2 应该很小；反之若 W^2 相当大，则应否定 H_0. 故否定域应取为：$W_0=\{W^2>\lambda\}$.

经计算知道，W^2 有下列简单公式：

$$W^2=\frac{1}{12n}+\sum_{i=1}^{n}\left[F_0(x_{(i)})-\frac{2i-1}{2n}\right]^2$$

从这个公式看出：如果 $F_0(x)$连续，H_0 成立，则 W^2 的分布与 $F_0(x)$无关. 当 n 很大时，也可找出 W^2 的近似分布，因而也可找出临界值 λ.

与柯氏检验法类似，我们顺便介绍一下 Smirnov 的两样本检验法.

设 $X_1,\cdots,X_{n_1}$ 是来自分布函数为 $F_1(x)$的样本，$Y_1,\cdots,Y_{n_2}$ 是来自分布函数为 $F_2(x)$的样本，且两个样本相互独立. 设 F_1,F_2 都是未知的连续函数. 研究检验问题：

$$H_0:F_1(x)\equiv F_2(x)\leftrightarrow H_a:F_1(x)\not\equiv F_2(x)$$

设 $F_{n_1}(x)$是 $X_1,\cdots,X_{n_1}$ 的经验分布函数，$G_{n_2}(x)$是 $Y_1,\cdots,Y_{n_2}$ 的经验分布函数. 令

$$D_{n_1,n_2}=\sup_x|F_{n_1}(x)-G_{n_2}(x)|$$

可以证明：在 H_0 成立的条件下，

$$\lim_{\substack{n_1\to\infty\\n_2\to\infty}}P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D_{n_1,n_2}\leqslant x\right)=Q(x)$$

其中 $Q(x)$的定义见(7.8).

给定 $\alpha\in(0,1)$，取否定域

$$W_0=\{D_{n_1,n_2}>\lambda\}$$

其中 λ 可查表得到.

(四) 正态性检验

在实际工作中常常需要检验一个随机变量是否服从正态分布，这叫做正态性检验. 上面介绍的 χ^2 检验法、柯氏检验法等当然

可以使用，但是上述方法由于是通用的，故有针对性不强的缺点. 对一些特定的分布往往可以找到针对这类特定分布功效较高的检验. 正态性检验的方法有很多种，经过国际标准化组织(ISO)统计标准分委员会的研究，最后认定 Wilk－Shapiro 的 W 检验和 D'Agostino的 D 检验是最好的，它们犯第二类错误的概率最小. 该委员会向世界各国推荐使用这两个正态性检验方法. 我国已把这两个方法列入国家标准，编号是 GB 488—85. 下面介绍这两个检验法. 当样本量 $n \leqslant 50$ 时用 W 检验，当样本量 $n>50$ 时用 D 检验.

(1) W 检验($n \leqslant 50$)

对于检验问题

$$H_0: X \text{ 服从正态分布} \leftrightarrow H_a: X \text{ 不服从正态分布}$$

取统计量

$$W=\frac{\left[\sum_{k=1}^{\left[\frac{n}{2}\right]} a_k\left(X_{(n+1-k)}-X_{(k)}\right)\right]^2}{\sum_{k=1}^{n}\left(X_{(k)}-\overline{X}\right)^2} \tag{7.9}$$

其中 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 是样本 $X_1, \cdots, X_n$ 的次序统计量，$\{a_k\}$ 是一组特定的值，可从特制的表中查得.

W 检验的否定域是

$$W_0=\{W<\lambda\} \tag{7.10}$$

其中 λ 满足 $P(W<\lambda \mid H_0)=\alpha$.

当样本量 $n \leqslant 50$ 时，国标附有相应的 $\{a_k\}$ 的值以及 $\alpha=0.01$, $0.05, 0.10$ 时 λ 的值.

这一检验法的理由大致如下：

设 $X_1, \cdots, X_n$ 独立同分布，均服从 $N(\mu, \sigma^2)$，则 $\frac{X_1-\mu}{\sigma}, \cdots$, $\frac{X_n-\mu}{\sigma}$ 独立同分布，均服从 $N(0,1)$. 令 $Y_i=\frac{X_{(i)}-\mu}{\sigma}$, $\varepsilon_i=X_{(i)}-EX_{(i)}$, $m_i=EY_i$，注意 $m_1, \cdots, m_n$ 是确定的数，与 μ, σ 无关. 显然

$X_{(i)}=\mu+\sigma m_i+\varepsilon_i$. 注意 n 较大时，$X_{(i)}\approx EX_{(i)}$，因而 ε_i 很小. 这表明 n 个点 $(X_{(i)},m_i)$ $(i=1,2,\cdots,n)$ 近似在一条直线上.

怎样判别 n 个点是否近似在一直线上呢？可用下列统计量：

$$R^2=\left[\frac{\sum_{i=1}^{n}(X_{(i)}-\overline{X})(m_i-\overline{m})}{\sqrt{\sum_{i=1}^{n}(X_{(i)}-\overline{X})^2\sum_{i=1}^{n}(m_i-\overline{m})^2}}\right]^2\quad\left(\overline{m}=\frac{1}{n}\sum_{i=1}^{n}m_i\right)$$

在 H_0 下，即诸 X_i 服从 $N(\mu,\sigma^2)$ 时 R^2 接近 1(见第四章). 可见 $R^2<\lambda$(λ 远小于 1)时则应拒绝 H_0.

由于 $N(0,1)$ 是对称分布，所以 $(Y_1,\cdots,Y_n)$ 与 $(-Y_n,\cdots,-Y_1)$ 有相同的联合分布，从而 Y_k 与 $-Y_{n+1-k}$ 同分布，故 $m_k=-m_{n+1-k}$ $(k=1,\cdots,n)$，$\overline{m}=0$，于是

$$R^2=\frac{\left(\sum_{i=1}^{n}X_{(i)}m_i\right)^2}{\sum_{i=1}^{n}(X_{(i)}-\overline{X})^2\sum_{i=1}^{n}m_i^2}=\frac{\left[\sum_{k=1}^{\left[\frac{n}{2}\right]}m_{n-k+1}(X_{(n+1-k)}-X_{(k)})\right]^2}{\sum_{i=1}^{n}(X_{(i)}-\overline{X})^2\sum_{i=1}^{n}m_i^2}$$

令

$$a_k=\frac{m_{n+1-k}}{\left(\sum_{i=1}^{n}m_i^2\right)^{\frac{1}{2}}}$$

即得 $W=R^2$.

可见，采用否定域 $W_0=\{W<\lambda\}$ 是有道理的.

当 $n>50$ 时，采用下列 D'Agostino 检验(D 检验)更好些.

(2) D 检验($n>50$)

对于检验问题：

$$H_0: X \text{ 服从正态分布} \leftrightarrow H_a: X \text{ 不服从正态分布}$$

取统计量

$$D = \frac{\sum_{k=1}^{n}\left(k - \frac{n+1}{2}\right)X_{(k)}}{(\sqrt{n})^3\sqrt{\sum_{k=1}^{n}(X_k - \overline{X})^2}}$$

$$Y = \frac{\sqrt{n}(D - 0.282\,094\,79)}{0.029\,985\,98}$$

在 H_0 下，Y 有确定的极限分布，记分布密度为 $g(u)$，取否定域 $W_0 = \{Y < \lambda_1 \text{ 或 } Y > \lambda_2\}$，其中 λ_1, λ_2 满足

$$\int_{\lambda_2}^{\infty} g(u)\mathrm{d}u = \frac{\alpha}{2}, \int_{\lambda_1}^{\infty} g(u)\mathrm{d}u = 1 - \frac{\alpha}{2}$$

国标附有 $\alpha = 0.01, 0.05, 0.10$ 时 λ_1, λ_2 的值.

§8 几种常用的非参数检验

在假设检验中，除了已知总体分布类型，对若干个未知参数作统计检验的参数统计检验外，还有一类所谓非参数检验. 非参数检验往往不假定总体的分布类型，直接对总体的分布的某种假设(例如对称性、分位数大小等等假设)作统计检验. 当然，上一节介绍的拟合优度检验也是非参数检验. 除了拟合优度检验外，还有许多常用的非参数检验. 最常见的非参数检验统计量有 3 类：计数统计量、秩统计量、符号秩统计量. 分别论述如下.

(一) 计数统计量

定义 8.1 设 X 是随机变量，对给定的实数 θ_0，记

$$\overline{\psi} = \psi(X - \theta_0)$$

其中

$$\psi(t) = \begin{cases} 1, & \text{当 } t > 0 \\ 0, & \text{当 } t \leqslant 0 \end{cases} \tag{8.1}$$

则称 $\overline{\psi}$ 为 X 按 θ_0 分段的计数统计量.

定理 8.1 设 $X_1, \cdots, X_n$ 相互独立，X_i 有连续分布函数 $F_i(x)$，又 $F_i(\theta_0) = p_0$，$0 < p_0 < 1$，$i = 1, \cdots, n$(即 $F_1(x), \cdots, F_n(x)$

有共同的 p_0 分位数 θ_0). $\bar{\psi}_1,\cdots,\bar{\psi}_n$ 分别是 $X_1,\cdots,X_n$ 的按 θ_0 分段的计数统计量. 则 $\bar{\psi}_1,\cdots,\bar{\psi}_n$ 相互独立同分布,共同分布是参数为 $1-p_0$ 的伯努利分布.

证明:显然.

从定理 8.1 知,任一依赖 $\bar{\psi}_1,\cdots,\bar{\psi}_n$ 的统计量 $T(\bar{\psi}_1,\cdots,\bar{\psi}_n)$对一切 p_0 分位数为 θ_0 的连续分布有固定不变的分布. 因而可用这样的统计量去检验总体分布的分位数的假设.

例 8.1 符号检验

设 $X_1,\cdots,X_n$ 是来自连续分布总体的简单随机样本. 对于检验问题:

$$H_0:F(0)=\frac{1}{2}\leftrightarrow H_1:F(0)\neq\frac{1}{2}$$

可采用检验统计量 $B=\sum_{i=1}^{n}\psi(X_i)$ (ψ 的定义见(8.1)),即 $X_1,\cdots,X_n$ 中取正号的个数. 在 H_0 下 B 的分布是参数为 n 和$\frac{1}{2}$的二项分布 $b\left(n,\frac{1}{2}\right)$. 给定显著性水平 α,可确定整数 C_1,C_2,(C_1 尽可能大,C_2 尽可能小)满足

$$P(B\leqslant C_1)\leqslant\frac{\alpha}{2},\quad P(B\geqslant C_2)\leqslant\frac{\alpha}{2}$$

取 $W=\{0,1,\cdots,C_1\}\cup\{C_2,C_2+1,\cdots,n\}$作为否定域,当且仅当 $B\in W$时拒绝 H_0.

同样地,若要检验 $H_0:F(\theta_0)=p_0$,只需取统计量 $B=\sum_{i=1}^{n}\psi(X_i-\theta_0)$,在 H_0 下它有分布 $b(n,1-p_0)$,由此就可确定检验的否定域.

符号检验是实际工作中一种常用的检验. 例如,在进行某项对比实验时,X 是未施加某一处理①的记录(某一特性值的测量

① "处理"(treatment)一词含义很广,可以指某项设计,可以指某项工艺,可以指某项手段,也可以指使用某种药品等等. "处理"可理解为"措施".

值)，Y 是施加某一处理的记录. 要判断有无处理效应(即判断施加处理与不施加处理在结果上有无差别)，常研究随机变量 $Y-X$. 若无处理效应，则 Y 与 X 有相同分布，$Y-X$ 的分布 $F(t)$ 是关于 0 对称的，因而 $F(0)=\frac{1}{2}$. 以 $Y_1-X_1,\cdots,Y_n-X_n$ 为样本，考虑检验问题：

$$H_0:F(0)=\frac{1}{2}\leftrightarrow H_1:F(0)\neq\frac{1}{2}$$

对此可用符号检验. 若否定 H_0，则判断有处理效应.

例 8.2 某工厂为提高某种产品的质量，对生产工艺进行了改变. 为了检验新工艺下的产品是否比原工艺下的产品质量确有改进，随机地取一新产品和一旧产品作为一对进行比较，共比较了 20 对. 记录如下("＋"表示新产品好，"－"表示旧产品好)：

＋，－，－，＋，＋，＋，＋，＋，－，＋，＋，＋，－
＋，＋，＋，＋，＋，＋，＋

检验问题是

H_0：新旧产品质量一样$\leftrightarrow H_1$：新产品比旧产品好

对此可用符号检验作单边检验. 用 B 表示观测结果中"＋"号的个数. 给定显著性水平 α，找尽可能小的整数 C 满足：$P(B\geqslant C|H_0)\leqslant\alpha$. 取 $W=\{C,C+1,\cdots,20\}$ 作为否定域，当且仅当 $B\in W$ 时拒绝 H_0. 若 $\alpha=0.05$，查本书附表 8 知 $C=15$. 由本例的实际结果算出 $B=16>15$，故否定 H_0，即认为新产品比旧产品在质量上确有改进.

(二) 秩统计量

定义 8.2 设 $Z_1,\cdots,Z_N$ 是来自连续分布 $F(z)$ 的样本，其次序统计量记为 $Z_{(1)}\leqslant\cdots\leqslant Z_{(N)}$，令

$$R_i=\min\{k:1\leqslant k\leqslant N,Z_i=Z_{(k)}\}\quad(i=1,2,\cdots,N)$$

称 R_i 为 Z_i 的秩.

定理 8.2 设 $Z_1,\cdots,Z_N$ 是来自连续分布 $F(z)$ 的样本，$R_1,\cdots,R_N$ 分别是 $Z_1,\cdots,Z_N$ 的秩，则随机向量 $R=(R_1,\cdots,R_N)$ 在集合

$$A=\{\gamma:\gamma\text{是}1,\cdots,N\text{的一个排列}\}$$

上均匀分布，即对 $1,\cdots,N$ 的任一排列 γ 有

$$P(R=\gamma)=\frac{1}{N!} \tag{8.2}$$

证明：设 $\gamma=(\gamma_1,\cdots,\gamma_N)\in A$，则

$$\begin{aligned}P(R=\gamma)&=P(R_1=\gamma_1,\cdots,R_N=\gamma_N)\\&=P((Z_1,\cdots,Z_N)=(Z_{(\gamma_1)},\cdots,Z_{(\gamma_N)}))\\&=P((Z_{d_1},\cdots,Z_{d_N})=(Z_{(1)},\cdots,Z_{(N)}))\\&=P(Z_{d_1}<\cdots<Z_{d_N})\end{aligned}$$

其中 d_i 是数 i 在排列 γ 中的位置，即 $i=\gamma_{d_i}$. 又因 $(Z_1,\cdots,Z_N)$ 与 $(Z_{d_1},\cdots,Z_{d_N})$ 同分布，所以

$$P(R=\gamma)=P(Z_1<Z_2<\cdots<Z_N)$$

这对一切 $\gamma\in A$ 均成立，故(8.2)成立. 证毕.

由定理 8.2 知，仅依赖 R 的统计量 $T(R)$ 对任一连续分布总体有固定不变的分布. 这一类统计量称秩统计量，在非参数检验中有广泛的应用. 最常用的秩统计量是下列例 8.3 中的 Wilcoxon 秩和统计量.

例 8.3 在对照实验①中有两个总体：$F(x)$ 和 $G(x)$. $X_1,\cdots,X_m$ 是对照组的观测值，它们是来自 $F(x)$ 的样本. $Y_1,\cdots,Y_n$ 是处理组的观测值，它们是来自 $G(x)$ 的样本. 检验问题是

$$H_0:F(x)\equiv G(x)\leftrightarrow H_1:G(x)\equiv F(x-c)$$

(其中 $c>0$，c 是处理效应).

将 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 这 $m+n=N$ 个随机变量混在一起排序(从小到大排列)，产生对应的秩：

$$Q_1,\cdots,Q_m,R_1,\cdots,R_n$$

---

① 对照实验在工农业试验与医学试验工作中常用，它的含义是：将要观察的个体分为两组，第一组的个体不施加某处理，第二组的个体皆施加某处理. 试验目的是比较施加某处理与不施加某处理在效果上有无不同. 此时第一组叫做对照组，第二组叫做处理组.

Wilcoxon 秩和统计量为

$$W = \sum_{i=1}^{n} R_i$$

直观上可以看出，若 H_1 为真，则 W 的值应该比较大. 由定理 8.2 知，在 H_0 下 $Q_1,\cdots,Q_m,R_1,\cdots,R_n$ 的联合分布是完全确定的，故 W 有确定的分布. 给定显著性水平 α，可确定临界值 $W(\alpha;m,n)$. 当且仅当 $W\geqslant W(\alpha;m,n)$ 时拒绝 H_0. 本书附表列出了常用的临界值.

例 8.4 用两种材料 A 及 B 制造同一种产品. 今分别随机抽取若干个进行比较，按产品性能（例如强度）从劣到优排列如下：

$$B,B,A,B,B,A,A,B,A,A,A,A$$

（这里 B 表示材料 B 制成的产品，A 表示材料 A 制成的产品）问：两种材料对产品性能有无不同影响？

此问题可换成双边检验问题：

$$H_0: A \text{ 与 } B \text{ 两种材料效果一样} \leftrightarrow H_1: A \text{ 与 } B \text{ 效果不同}$$

用 Wilcoxon 秩和统计量作检验. W 等于 B 材料产品在排序中的秩之和. 给定显著性水平 α，确定临界值 $W_1\left(\frac{\alpha}{2};m,n\right)$ 和 $W_2\left(\frac{\alpha}{2};m,n\right)$ 满足

$$P\left(W < W_1\left(\frac{\alpha}{2};m,n\right) \middle| H_0\right) \leqslant \frac{\alpha}{2}$$

$$P\left(W > W_2\left(\frac{\alpha}{2};m,n\right) \middle| H_0\right) \leqslant \frac{\alpha}{2}$$

当且仅当 $W<W_1\left(\frac{\alpha}{2};m,n\right)$ 或 $W>W_2\left(\frac{\alpha}{2};m,n\right)$ 时拒绝 H_0. 本例中 $m=7,n=5$. 取 $\alpha=0.10$，查 W 的临界值表得 $W_1=22,W_2=43$. 由样本算得 $W=1+2+4+5+8=20<22$，故样本落入否定域. 我们应否定 H_0，认为 A,B 两种材料对产品质量有显著差异.

（三）符号秩统计量

定义 8.3 设 $X_1,\cdots,X_n$ 是来自连续分布的样本，随机变量

$|X_1|,\cdots,|X_n|$中$|X_i|$的秩称为 X_i 的绝对秩，记为 R_i^+. 而 $\bar{\psi}_i R_i^+$ 称为 X_i 的符号秩，其中 $\bar{\psi}_i=\psi(X_i)$是 X_i 按 0 分段的计数统计量.

定理 8.3 设 X 是随机变量，其分布函数连续且概率分布关于 0 对称，则 $\bar{\psi}=\psi(X)$与$|X|$独立.

证明：若 $t<0$，则 $P(\bar{\psi}=0,|X|\leqslant t)=0=P(\bar{\psi}=0)P(|X|\leqslant t)$，$P(\bar{\psi}=1,|X|\leqslant t)=0=P(\bar{\psi}=1)P(|X|\leqslant t)$；若 $t>0$，则

$$\begin{aligned}P(\bar{\psi}=1,|X|\leqslant t)&=P(X>0,|X|\leqslant t)\\&=P(0<X\leqslant t)\\&=\frac{1}{2}P(|X|\leqslant t).\\&=P(\bar{\psi}=1)P(|X|\leqslant t)\end{aligned}$$

类似地有

$$P(\bar{\psi}=0,|X|\leqslant t)=P(\bar{\psi}=0)P(|X|\leqslant t)$$

由独立性定义知 $\bar{\psi}$ 与$|X|$独立. 证毕.

定理 8.4 设 $X_1,\cdots,X_n$ 是关于 0 对称的连续分布的样本，则其对应的计数统计量 $\psi(X_1),\cdots,\psi(X_n)$与它们的绝对秩 $R^+=(R_1^+,\cdots,R_n^+)$相互独立. $\psi(X_i)$服从 $p=\frac{1}{2}$ 的伯努利分布($i=1,\cdots,n$)，R^+ 在 $A=\{\gamma:\gamma$ 为 $1,2,\cdots,n$ 的排列$\}$上均匀分布.

证明：由定理 8.3、定理 8.1、定理 8.2 立刻可得本定理的结论.

由定理 8.4 知，只依赖 $\bar{\psi}=(\psi(X_1),\cdots,\psi(X_n))$与 $R^+=(R_1^+,\cdots,R_n^+)$的统计量 $T(\bar{\psi},R^+)$对关于 0 对称的连续分布总体有固定不变的分布，因而可用这类统计量构造关于总体分布对称性的检验.

例 8.5 Wilcoxon 符号秩统计量

设 $X_1,\cdots,X_n$ 是来自 $F(x)$的样本，$F(x)$连续且有对称点 θ (即 $F(\theta-x)=1-F(\theta+x)$). 考虑检验问题：

$$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta>\theta_0$$

令 $Z_i=X_i-\theta_0 (i=1,\cdots,n)$，$Z_1,\cdots,Z_n$ 的符号秩记为 $\psi(Z_1)R_1^+,\cdots,\psi(Z_n)R_n^+$. Wilcoxon 符号秩统计量为

$$W^+=\sum_{i=1}^{n}\psi(Z_i)R_i^+$$

在 H_1 下 W^+ 应有较大的值. 在 H_0 下 W^+ 有确定的分布，故给定显著性水平 α，有确定的临界值 $W^+(\alpha;n)$ 满足

$$P(W^+\geqslant W^+(\alpha;n)\mid H_0)\leqslant\alpha$$

当且仅当 $W^+\geqslant W^+(\alpha;n)$ 时拒绝 H_0. $W^+(\alpha;n)$ 的值可从特制的表中查出.

在结束本章之前，我们给出下面一个很有启发性的例子，表明检验问题的提法和处理方法是灵活多样的.

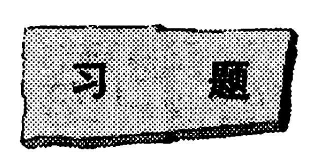
气象学家记录了 1942—1981 年间共 40 年 Raleigh(位于美国北卡罗来纳州)的年平均气温. 数据如表 8.1(单位:华氏温度).

表 8.1 Raleigh 年平均气温

年	温度	年	温度	年	温度	年	温度
1942	61.0	1952	60.3	1962	58.2	1972	58.5
1943	60.6	1953	61.3	1963	57.5	1973	59.8
1944	59.8	1954	60.1	1964	58.9	1974	59.7
1945	60.3	1955	59.6	1965	59.1	1975	59.6
1946	60.4	1956	59.9	1966	58.7	1976	58.7
1947	59.1	1957	60.1	1967	59.1	1977	58.6
1948	59.8	1958	57.4	1968	58.4	1978	59.1
1949	61.4	1959	59.9	1969	56.9	1979	58.6
1950	59.9	1960	57.9	1970	57.8	1980	59.3
1951	60.1	1961	58.7	1971	59.0	1981	58.3

以时间(年份)为 x 轴,温度值为 y 轴. 可在平面上画出 40 个点(所谓散点图). 由此可以近似看出,平均气温有下降的趋势. 现在问:如何判断 Raleigh 是否真的在变冷?

对此问题,不同的统计学家可能利用不同的方法进行检验,这是因为大家可能做出了不同的假定. 如果我们采取尽可能弱的假定,例如不假定温度服从任何特定的分布类型,则可以将 40 个数据看成一个随机的排列. 任取其中两个温度,可以构成一个"对". 更具体说,设第 i 年的温度是 y_i $(i=1942,1943,\cdots,1981)$. 称 $(y_i,y_j)(i<j)$ 是协调的,若 $y_i<y_j$;称 $(y_i,y_j)(i<j)$ 是不协调的,若 $y_i>y_j$. 设"协调对"的总数是 M_1,"不协调对"的总数是 M_2. $M=M_1-M_2$. 从直观上看,若 M_1 比 M_2 偏大则说明有变暖的趋势,若 M_1 比 M_2 偏小则说明有变冷的趋势. 若 M_1 与 M_2 相差不大,即 $|M|$ 靠近 0 时,则表明无变暖或变冷的趋势. 设 H_0 是假设:无趋势. 则据上面的分析,当 $|M|$ 较大时应拒绝 H_0. 这个检验法是合理的,但 $|M|$ 多大才应拒绝 H_0 呢? 即如何确定临界值? 注意 n 个不同年份的温度共有 $n!$ 种排列方式,在 H_0 下,它们应是等可能的. 因此,可以这样定义概率空间 $(\Omega,\mathscr{F},P)$,其中 Ω 由所有的排列组成(共 $n!$ 个元素),$\mathscr{F}$ 由 Ω 的所有子集组成,而 P 给予 Ω 的每个元素以 $\frac{1}{n!}$ 的概率. 有了概率空间,就可得到 H_0 下 M 的概率分布. 在本例中 $n=40$,计算较繁琐. 根据上面的 40 个数据,计算出 M 的数值是 -329. 但经过理论研究,在 H_0 下随机变量 M 不超过 -329 的概率小于 0.001. 因此在检验水平 0.001 下应拒绝 H_0. 换句话说,在 Raleigh 年平均气温有下降的趋势. (此例摘自 The fassination of statistics, Edited by R. J. Brook et al., p. 161—172, Marcel Dekker inc. 1986). 所用的检验就是所谓的 Mann trend test.



习 题

1. 设 X_1,X_2,X_3 是来自 Bernoulli $b(1,p)$ 分布的样本,检验问题

$$H_0: p=\frac{1}{2} \leftrightarrow H_a: p=\frac{3}{4}$$

的一个否定域为

$$W=\{(x_1,x_2,x_3): x_1+x_2+x_3 \geqslant 2\}$$

试求该否定域的第一、二类错误概率和 $p=\frac{3}{4}$时的功效.

2. 设 $X_1,X_2,\cdots,X_n$ 是来自密度函数为 $f(x)$的总体的样本，关于 $f(x)$ 有假设检验问题：

$$H_0: f(x)=f_0(x) \leftrightarrow H_a: f(x)=f_1(x)$$

其中 $f_0(x)$和 $f_1(x)$都是已知函数，令 W 为假设 H_0 的否定域，试通过$f_0(x)$ 和 $f_1(x)$表示两类错误概率.（$f_0(x)$不恒等于 $f_1(x)$).

3. 设 X 是$\left(\theta-\frac{1}{2},\theta+\frac{1}{2}\right)$上的均匀分布的一个观测值，考虑假设检验问题

$$H_0: \theta \leqslant 3 \leftrightarrow H_a: \theta \geqslant 4$$

构造检验法，使其功效函数 $\rho(\theta)$满足

$$\rho(\theta)=0，当\ \theta \leqslant 3；\rho(\theta)=1，当\ \theta \geqslant 4$$

（即两类错误概率均为零）

4. 设随机变量 X 的分布密度 $f(x)$可取下面的 $f_0(x)$或 $f_1(x)$：

$$f_0(x)=\begin{cases}1, & 0 \leqslant x \leqslant 1 \\ 0, & 其他\end{cases}$$

$$f_1(x)=\begin{cases}2x, & 0 \leqslant x \leqslant 1 \\ 0, & 其他\end{cases}$$

基于一个观测 X，来回答检验问题：

$$H_0: f(x)=f_0(x) \leftrightarrow H_a: f(x)=f_1(x)$$

取检验水平 0.1，求出使第二类错误概率 β 最小的检验法.

5. 设随机变量 X 有分布密度 $f(x;\theta)$，θ 未知，$x_1,x_2,\cdots,x_n$ 为 X 的样本，考虑假设检验问题：

$$H_0: \theta=\theta_0 \leftrightarrow H_a: \theta=\theta_1$$

其中 $\theta_0 \neq \theta_1$ 是给定值. 设 a,b 为给定的正常数. 记

$$L(\underline{x},\theta)=\prod_{i=1}^{n} f(x_i,\theta) \quad (\underline{x}=(x_1,\cdots,x_n))$$

考虑如下检验法 δ^*：当 $aL(x,\theta_0)>bL(\underline{x},\theta_1)$时接受 H_0，当 $aL(\underline{x},\theta_0)<$

$bL(\underset{\sim}{x},\theta_1)$时拒绝 H_0，如果 $aL(x,\theta_0)=bL(\underset{\sim}{x},\theta_1)$则既可接受 H_0 也可拒绝 H_0. 设 α^*,β^* 为 δ^* 的第一、二类错误概率，试证对任意检验法 δ，若 α,β 为 δ 的第一、二类错误概率，必有

$$a\alpha^* + b\beta^* \leqslant a\alpha + b\beta$$

6. 设 $X\sim N(\mu_0,\sigma^2)$，μ_0 已知，$X_1,X_2,\cdots,X_n$ 是 X 的样本. 试分别求出下列检验问题

(1) $H_0:\sigma^2\leqslant\sigma_0^2\leftrightarrow H_a:\sigma^2>\sigma_0^2$（$\sigma_0$ 已知）

(2) $H_0:\sigma^2\leqslant\sigma_1^2$ 或 $\sigma^2\geqslant\sigma_2^2\leftrightarrow H_a:\sigma_1^2<\sigma^2<\sigma_2^2$

（σ_1,σ_2 已知，$\sigma_1<\sigma_2$）

的 UMP 检验（检验水平为 α，以下各题非特别指出者皆同此）.

7. 设 $X\sim N(\mu,1)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，试找出检验问题

$$H_0:\mu\leqslant\mu_1 \text{ 或 } \mu\geqslant\mu_2\leftrightarrow H_a:\mu_1<\mu<\mu_2$$

$$(\mu_1,\mu_2 \text{ 已知}, \mu_1<\mu_2)$$

的 UMP 检验.

8. 设 $X_1,X_2,\cdots,X_n$ 为参数 λ 未知的 Poisson 分布的样本，试选适当的检验水平 α 求检验问题

$$H_0:\lambda=\lambda_0\leftrightarrow H_a:\lambda>\lambda_0 \quad (\lambda_0 \text{ 已知})$$

的 UMP 检验.

9. 设 $X_1,X_2,\cdots,X_n$ 为参数 p 未知的 Bernoulli 分布 $b(1,p)$的样本，试选适当的检验水平 α 求检验问题

$$H_0:p=p_0\leftrightarrow H_a:p\neq p_0 \quad (p_0 \text{ 已知})$$

的一致最大功效无偏检验（UMPU 检验）.

10. 设 $X_1,X_2,\cdots,X_n$ 是$[0,\theta]$上的均匀分布的样本，求假设检验问题

$$H_0:\theta=\theta_0\leftrightarrow H_a:\theta>\theta_0$$

的 UMP 检验.

11. 设 $X_1,X_2,\cdots,X_n$ 是 $N(\mu,1)$的样本，对检验问题

$$H_0:\mu\leqslant 0\leftrightarrow H_a:\mu>0$$

(1) 求出 $\alpha=0.025$ 的 UMP 检验，并求出其功效函数 $\rho(\mu)$.

(2) 为了使 $\mu\geqslant0.5$ 时上述检验的功效 $\rho(\mu)\geqslant0.9$，样本量 n 至少应取多大?

(3) 为了使 $\mu\leqslant-0.1$ 时 $\rho(\mu)\leqslant0.001$，样本量 n 至少应取多大?

12. 设 $X\sim N(0,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，σ_0 是已知正数.

(1) 对检验问题

$$H_0:\sigma=\sigma_0 \leftrightarrow H_a:\sigma>\sigma_0$$

找出 UMP 检验法;

(2) 对检验问题

$$H_0:\sigma=\sigma_0 \leftrightarrow H_a:\sigma<\sigma_0$$

找出 UMP 检验法;

(3) 证明对于检验问题

$$H_0:\sigma=\sigma_0 \leftrightarrow H_a:\sigma\neq\sigma_0$$

不存在一致最大功效检验.

13. 由经验知某零件重量 $X\sim N(\mu,\sigma^2)$, $\mu=15$, $\sigma^2=0.05$, 技术革新后, 抽了 6 个样品, 测得重量(单位:g)为

14.7, 15.1, 14.8, 15.0, 15.2, 14.6

已知方差不变, 问平均重量是否仍为 15? ($\alpha=0.05$)

14. 设 $f_n(x)$ 是 n 个自由度的 t 分布的密度, 即

$$f_n(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

试证明

$$\lim_{n\to\infty} f_n(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\quad (对一切\ x\in(-\infty,+\infty))$$

15. 设 X_1,X_2,X_3,X_4 为正态分布 $N(\mu,1)$ 的样本, 对假设检验问题

$$H_0:\mu\geqslant 10 \leftrightarrow H_a:\mu<10$$

求出水平 $\alpha=0.1$ 的 UMP 检验, 并求 $\mu=9$ 时此检验法的功效, 求出 $\mu=11$ 时接受 H_0 的概率.

16. 从某正态总体中抽取了 9 个数据, 得 $\overline{X}=0.4$,

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}=1.0,$$

在水平 0.05 和 0.01 下分别检验

(1) $H_0:\mu\leqslant 0 \leftrightarrow H_a:\mu>0$

(2) $H_0:\mu\geqslant 0 \leftrightarrow H_a:\mu<0$

解释结果的意义. 如果样本量增加到了 25, 而统计量 $\overline{X}$ 与 S 仍分别为 0.4 和 1.0, 再进行上面的检验并说明结果的意义.

17. 某糖厂用自动打包机打包, 每包重量的额定值为 100 kg, 每天开工

后需要检验一下打包机工作是否正常，即检查打包机是否有系统偏差. 某日开工后测得 9 包的重量(单位:kg)为

99.3,98.7,100.5,101.2,98.3,99.7,99.5,102.1,100.5

问:该日打包机工作是否正常?($\alpha=0.05$;已知包重服从正态分布)

18. 正常人的脉搏平均为 72 次/分，某医生测得 10 例慢性四乙基铅中毒患者的脉搏(次/分)如下:

54,67,68,78,70,66,67,70,65,69

已知这些患者的脉搏服从正态分布，问:四乙基铅中毒患者的脉搏与正常人的脉搏有无显著差异?($\alpha=0.05$)

19. 用热敏电阻测温仪间接测量地热勘探井底的温度，重复测量 7 次，测得温度(℃)为

112.0,113.4,111.2,112.0,114.5,112.9,113.6

而用某种精确方法测得的温度为 112.6(可看作温度真值)，试问:用热敏电阻测温仪间接测温有无系统偏差?($\alpha=0.05$)

20. 某工厂采用新法处理废水，对处理后的水测量所含某种有毒物质的浓度，得到 10 个数据(单位:mg/L):

22,14,17,13,21,16,15,16,19,18

而以往用老法处理废水后，该种有毒物质的平均浓度为 19. 问:从对这种有毒物质的处理来看，新法是否比老法效果要好?($\alpha=0.05$)

21. 有某种导线，要求其电阻的标准差不得超过 0.005(Ω). 今在生产的一批导线中取样品 9 根，测得 $S=0.007$(Ω)，设总体为正态分布，问在 0.05 水平下能认为这批导线的标准差显著地偏大吗?

22. 某机床厂某日从两台机器所加工的同一种零件中，分别抽若干个样测量零件尺寸，得

第一台机器的 6.2,5.7,6.5,6.0,6.3,5.8,5.7,6.0,6.0,5.8,6.0

第二台机器的 5.6,5.9,5.6,5.7,5.8,6.0,5.5,5.7,5.5

问:这两台机器的加工精度是否有显著性差异?($\alpha=0.05$)

23. 检查了 26 匹马，测得每 100 mL 的血清中，所含无机磷平均为 3.29 mg，标准差为 0.27 mg;又检查了 18 头羊，100 mL 的血清中含无机磷平均为 3.96 mg，标准差为 0.40 mg. 试以 0.05 的检验水平，检验马与羊的血清中含无机磷的量是否有显著性差异?

24. 10 个失眠患者，服用甲、乙两种安眠药，延长睡眠时间(h)如下表

所示：

患者编号	1	2	3	4	5	6
甲安眠药	1.9	0.8	1.1	0.1	−0.1	4.4
乙安眠药	0.7	−1.6	−0.2	−1.2	−0.1	3.4
患者编号	7	8	9	10		
甲安眠药	5.5	1.6	4.6	3.4		
乙安眠药	3.7	0.8	0	2.0		

问：这两种安眠药的疗效有无显著性差异？（可以认为服用两种安眠药后增加的睡眠时间之差近似服从正态分布，$\alpha=0.05$）

25. 比较甲、乙两种安眠药的疗效. 将 20 个患者分成两组，每组 10 人；甲组病人服用甲种安眠药，乙组病人服用乙种安眠药，如服药后延长的睡眠时间分别近似服从正态分布，其数据仍如上题（这时数据不再有成对的关系），问这两种安眠药的疗效有无显著性差异？（$\alpha=0.05$）

26. 设 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，$Y_1,Y_2,\cdots,Y_m$ 是 Y 的样本. 试用广义似然比方法导出下列检验问题的具体否定域：

$$H_0:\mu_1=\mu_2\leftrightarrow H_a:\mu_1\neq\mu_2 \quad (\text{已知 } \sigma_1^2=\sigma_2^2)$$

27. 设 $X_1,X_2,\cdots,X_n$ 是来自指数分布

$$f(x,\theta_1)=\frac{1}{\theta_1}e^{-\frac{x}{\theta_1}}$$

$(x\geqslant 0)$的样本，$Y_1,Y_2,\cdots,Y_m$ 是来自指数分布

$$f(x,\theta_2)=\frac{1}{\theta_2}e^{-\frac{x}{\theta_2}}$$

$(x\geqslant 0)$的样本，且两个样本相互独立. 这里 θ_1,θ_2 是未知的正参数.

（1）对检验问题

$$H_0:\theta_1=\theta_2\leftrightarrow H_a:\theta_1\neq\theta_2$$

找出广义似然比检验法.

（2）证明上述检验法的否定域只依赖于比值$\dfrac{\sum_{i=1}^{n}X_i}{\sum_{i=1}^{m}Y_i}$.

(3) 你能求出统计量$\dfrac{\sum_{i=1}^{n}X_i}{\sum_{i=1}^{m}Y_i}$在 H_0 下的分布吗?

28. 在一正 20 面体的 20 个面上,分别标以数字 0,1,2,…,9,每个数字在两个面上标出.为检验它是否质地匀称,共做了 800 次投掷试验,数字 0,1,2,…,9 朝正上方的次数如下:

数字	0	1	2	3	4	5	6	7	8	9
频数	74	92	83	79	80	73	77	75	76	91

问:该 20 面体是否匀称?

29. 试用柯氏检验法检验下列 25 个数据是否来自 $N(0,1)$分布($\alpha=0.05$):

−2.46 −2.11 −1.23 −0.99 −0.42 −0.39 −2.21
−0.15 −0.10 −0.07 −0.02 0.27 0.40 0.42 0.44
0.70 0.81 0.88 1.07 1.39 1.40 1.47 1.62 1.64 1.76

30. 在对 UFO 的调查中,查问了 1 276 个自称看到飞碟的人,并按他们看到飞碟的地点及飞碟的形态进行分类,结果如下:

		看到飞碟的地点	
		在西班牙	不在西班牙
看到的飞碟的形态	停在地上的碟形	53	705
	飞翔的碟形	38	412
	其他形式	9	59

检验“看到飞碟的地点”与“看到的飞碟的形态”是否独立($\alpha=0.05$),并说明检验的结果是支持还是反对飞碟的存在性.

31. 甲乙二人分析同一气体的 CO_2 含量,当二人的分析数据没有个人间的差异时,应有相同分布,从而“甲的数据—乙的数据”的分布 $F(x)$关于 0 对称. 可用符号检验法检验 $H_0:F(0)=\frac{1}{2}$.若否定 H_0,则断定甲乙二人的分析结果有显著差异. 试用下列实测数据作上述检验(检验水平 $\alpha=0.05$):

甲	14.7	15.0	15.2	14.8	15.5	14.6	14.9	14.8	15.1
乙	14.6	15.1	15.4	14.7	15.2	14.7	14.8	14.6	15.2
甲	14.7	14.8	14.7	15.0	14.9	14.9	15.2	14.7	15.4
乙	14.6	14.6	14.8	15.3	14.7	14.6	14.8	14.9	15.2

32. 用两种材料的灯丝制造灯泡，设甲材料灯泡的寿命分布为 $F(x)$，乙材料灯泡的寿命分布为 $F(x-\delta)$. 今分别就两种材料灯泡随机抽取若干个进行寿命试验，得到寿命（单位:h）数据如下：

甲材料　1 610，1 650，1 680，1 710，1 750，1 720，1 800

乙材料　1 580，1 600，1 640，1 630，1 700

试用秩和检验法检验两种材料对灯泡寿命的影响有无显著差异？（检验水平 $\alpha=0.05$）

33. 设两样本 $X_1,\cdots,X_m;Y_1,\cdots,Y_n$ 一起排序得到的秩统计量为 $(Q_1,Q_2,\cdots,Q_m;R_1,R_2,\cdots,R_n)$，Wilcoxon 秩和统计量 $W=\sum_{i=1}^{n}R_i$，试证明

$$W=\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(Y_j-X_i)+\frac{1}{2}n(n+1)$$

其中

$$\psi(t)=\begin{cases}1, & t>0\\ 0, & t\leqslant 0\end{cases}$$

第四章　回归分析与线性模型

§1　引　　言

回归分析是数理统计学的一个重要组成部分,它的任务是研究变量之间的相互关系,建立变量之间的经验公式,以便达到预测和控制的目的.

线性模型是数理统计学中一种应用十分广泛的模型,线性回归分析、方差分析、试验设计后的数据分析等都可以归为这种模型的统计分析.

(一) 变量之间的相互关系

变量之间如果不是毫无联系,那么它们之间的关系可以分为两种类型:确定性关系(或称函数关系)与相关关系.

确定性关系是某些变量值确定以后另一些变量的值就可以完全被确定的关系. 比如在物理学中,在一个阻值为 R 的电阻上加电压 U,则电流一定是 $I=\frac{U}{R}$. 这里变量 I 的值被变量 U 和 R 的值完全确定,I 与 U 和 R 的关系就是确定性关系.

相关关系是指一些变量的值确定以后,另一些变量的值并不能完全确定,但与前者有某种联系. 比如一个人的身高与体重的关系. 一个人身高越高,体重一般也越重,但身高与体重之间并不是严格的函数关系,有些人个子不高,但很重,也有人个子很高,但体重不太重. 如果记身高为 x,体重为 y,则我们可以用如下的数学表达式表示它们之间的关系:

$$y=a+bx+e$$

其中 e 是一个随机项，此式说明了身高越高体重也应越大，但可能有随机摆动 e. 类似地，一个变量 y 对多个变量 $x_1, x_2, \cdots, x_p$ 的相关关系一般写作

$$y=f(x_1, x_2, \cdots, x_p)+e$$

这里 f 是确定性函数，e 是期望为零的随机变量.

（二）回归分析

统计学中回归分析的任务就是要研究变量 y 与变量 $x_1, x_2, \cdots, x_p$ 之间的关系①.

（1）是否有相关关系. 即 $x_1, x_2, \cdots, x_p$ 的值定下以后，y 的值是否在某个确定的范围内.

（2）若有相关关系，如何找出适当的函数 $f(x_1, x_2, \cdots, x_p)$ 使得

$$y=f(x_1, x_2, \cdots, x_p)+e \tag{1.1}$$

其中 e 在某种意义下很小. 这个 e 叫做随机项或随机误差项，它是期望为零的随机变量（如果 $Ee\neq 0$，我们可以把（1.1）式右边改写成 $[f(x_1, x_2, \cdots, x_p)+Ee]+(e-Ee)$，以 $e-Ee$ 作为新的随机误差项）.

（3）如果得到的关系式（1.1）中 e 的方差很小，则有

$$y\approx f(x_1, x_2, \cdots, x_p) \tag{1.2}$$

称此式为回归关系式或经验公式，它可以用于预测和控制.

① 预测：已知 $x_1, x_2, \cdots, x_p$ 取值为 $x_1^0, x_2^0, \cdots, x_p^0$，求相应的 y 大约是多少. 比如通过分析知道人的身高与体重之间有回归关系式

$$y\approx a+bx$$

现知某人身高为 x^0，那么我们就知道此人体重大约为 $a+bx^0$.

② 控制：如果变量 $x_1, x_2, \cdots, x_p$ 是可以控制取值的，而 y 不

---

① 数理统计学中对多个变量 $y_1, y_2, \cdots, y_m$ 与多个变量 $x_1, x_2, \cdots, x_n$ 的相关关系也有研究，一般列入多元统计分析的范围.

能直接控制，y 与 $x_1, x_2, \cdots, x_p$ 之间有回归关系式(1.2)，那么如果要控制 y 在区间(A,B)内则可利用(1.2)的关系通过控制 $x_1, x_2, \cdots, x_p$ 来实现. 比如通过分析知道，在对某种产品表面进行腐蚀刻线时，腐蚀时间 x 与腐蚀深度 y 之间有回归关系式

$$y \approx a + bx \tag{1.3}$$

现在要控制腐蚀深度 y 在区间(A,B)内，则可通过控制腐蚀时间 x 于区间(C,D)内实现.

预测、控制问题对生产实践、科学实验、管理工作等许多方面都有重要的应用价值.

怎样研究变量间的相关关系呢？从物理学、化学、生物学、经济学、社会学等专业的理论常常只能得到一些定性的看法，有时连定性的看法也得不到. 数理统计学的基本手段是从数据出发研究这些问题，换句话说，是通过调查或进行科学实验获得大量数据并在此基础上进行科学的分析以回答上述要研究的问题.

回归分析需要的数据可以列成一个表，见表 1.1.

表 1.1

观测号 \ 观测值 \ 变量	x_1	x_2		x_p	y
1	x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
2	x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\vdots$	$\vdots$			$\vdots$	
n	x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

其中 y_i 是 $x_1 = x_{i1}, x_2 = x_{i2}, \cdots, x_p = x_{ip}$ $(i = 1, 2, \cdots, n)$时的 y 的观测值，一般要求观测值数 n 大于自变量的个数 p.

(三) 回归分析的两种观点

回归分析的出发点是表 1.1 那样的数据. 这里变量 y 一般

是不能人为控制其取值的，只能通过观测得到，回归分析中把它作为因变量．变量 $x_1,x_2,\cdots,x_p$ 有些情况下可以人为控制取值，有些情况下也只能观测得到．由此就产生了回归分析中的两种观点：

(1) 把自变量 $x_1,x_2,\cdots,x_p$ 看成非随机的普通变量，因变量 y 看成随机变量，这称为第一类回归分析．如果 $x_1,x_2,\cdots,x_p$ 的取值是人为控制的，则这种观点比较自然；如果 $x_1,x_2,\cdots,x_p$ 的值是观测得到的，因得到的数据仍与可以控制时同样具有表 1.1 那样的形式，所以这时也完全可以用第一类回归来处理．

(2) 把 $x_1,x_2,\cdots,x_p$ 与 y 都看成随机变量，这称为第二类回归分析．如果 $x_1,x_2,\cdots,x_p$ 是观测得到的，则这种观点比较自然，但实际工作中人们一般还是用第一类回归分析来处理．

我们将主要讨论第一类回归分析，关于第二类回归分析只作简单的讨论．我们将指出，在许多情形下，两类回归分析有相同的计算公式．

(四) 线性回归分析与线性模型

回归分析的基本问题是：如何从表 1.1 那样的数据出发找出(1.1)式中的函数 f 使得(1.1)中的随机项 e 在某种意义下最小？

函数 f 的可选范围太广了，难以下手．如果预先假定 f 是线性函数：

$$f(x_1,x_2,\cdots,x_p)=b_0+b_1x_1+\cdots+b_px_p$$

($b_0,b_1,\cdots,b_p$ 均未知)，则模型(1.1)变成

$$y=b_0+b_1x_1+\cdots+b_px_p+e \tag{1.4}$$

称之为线性回归模型．结合表 1.1 的数据可得如下关系式：

$$\begin{aligned}
y_1&=b_0+b_1x_{11}+b_2x_{12}+\cdots+b_px_{1p}+e_1\\
y_2&=b_0+b_1x_{21}+b_2x_{22}+\cdots+b_px_{2p}+e_2\\
&\cdots\cdots\cdots\cdots\\
y_n&=b_0+b_1x_{n1}+b_2x_{n2}+\cdots+b_px_{np}+e_n
\end{aligned} \tag{1.5}$$

称之为线性模型.

线性回归分析的基本问题就是如何确定 $b_0, b_1, \cdots, b_p$ 使得(1.4)中的 e 在某种意义下最小.

线性函数是极特殊的多元函数,但线性回归分析却是回归分析里最重要的组成部分. 这是为什么呢? 原因有二:① 线性回归模型在数学上有成熟的处理方法,线性代数的工具可以发挥其强大的威力,这一点在本章中将充分表现出来. ② 实际当中不仅是经常遇到线性回归模型,而且许多非线性回归模型经过适当的变换可以化为线性回归模型. 这一点现作如下解释.

例 1.1 在彩色显影中,根据以往的经验,染料光学密度 y 与析出银的光学密度 x 之间有下面类型的关系

$$y \approx A\mathrm{e}^{-B/x} \ (B>0)$$

其中 A, B 未知. 这里 y 与 x 之间不是线性关系,但令 $y^* = \ln y$, $x^* = \dfrac{1}{x}$,则

$$y^* \approx \ln A - Bx^*$$

即 y^* 与 x^* 有近似的线性关系.

一般地,一元多项式回归模型常可化为多元线性回归模型. 如设

$$y = b_0 + b_1 x + \cdots + b_p x^p + e$$

则只要令 $x_j = x^j (j=1,2,\cdots,p)$,就有

$$y = b_0 + b_1 x_1 + \cdots + b_p x_p + e$$

即多元线性回归模型.

上述方法还可以推广. 设

$$y = f(x_1, x_2, \cdots, x_p) + e \tag{1.6}$$

其中 f 是连续函数,它可用多项式

$$\varphi(x_1, x_2, \cdots, x_p) = \sum_{\substack{i_1, i_2, \cdots, i_p \geqslant 0 \\ i_1 + i_2 + \cdots + i_p \leqslant m}} a(i_1, i_2, \cdots, i_p) x_1^{i_1} x_2^{i_2} \cdots x_p^{i_p}$$

来逼近. 记 $x_{i_1 i_2 \cdots i_p} = x_1^{i_1} x_2^{i_2} \cdots x_p^{i_p}$,则

$$\varphi(x_1, x_2, \cdots, x_p) = \sum_{\substack{i_1, i_2, \cdots, i_p \geqslant 0 \\ i_1 + i_2 + \cdots + i_p \leqslant m}} a(i_1, i_2, \cdots, i_p) x_{i_1 i_2 \cdots i_p}$$

是新变量 $\{x_{i_1 i_2 \cdots i_p}\}$ 的线性函数,于是可以用线性回归模型 $y = \varphi(x_1, x_2, \cdots, x_p) + \tilde{e}$ 来近似模型(1.6).

更一般地,对于模型

$$y = a_0 + \sum_{j=1}^{l} a_j \psi_j(x_1, x_2, \cdots, x_p) + e$$

只要 ψ_j 是已知函数,则令

$$z_j = \psi_j(x_1, x_2, \cdots, x_p) \quad (j = 1, 2, \cdots, l)$$

便有

$$y = a_0 + \sum_{j=1}^{l} a_j z_j + e$$

成为线性回归模型.

综上所述,线性回归模型的应用是十分广泛的. 为了便于理解,下面先介绍一元线性回归,然后系统讲述线性模型的一般理论,最后具体讨论线性回归问题.

§2 一元线性回归

(一) 问题的提出

在一元线性回归分析里,我们要考察的是:随机变量 y 与普通变量 x(y 和 x 也可写成 Y 和 X)之间的联系. 对于 x 和 y,通过观测或试验得到若干对数据

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

在此基础上,如何获得 y 对 x 的经验公式呢? 先看一个例子.

例 2.1 某种合成纤维的强度 y 与其拉伸倍数 x 有关. 表 2.1是 24 个纤维样品的强度与相应的拉伸倍数的实测记录. 希望找出强度 y 与拉伸倍数 x 的近似关系式.

表 2.1 合成纤维数据

编号	拉伸倍数 x	强度 y	编号	拉伸倍数 x	强度 y
1	1.9	1.4	13	5.0	5.5
2	2.0	1.3	14	5.2	5.0
3	2.1	1.8	15	6.0	5.5
4	2.5	2.5	16	6.3	6.4
5	2.7	2.8	17	6.5	6.0
6	2.7	2.5	18	7.1	5.3
7	3.5	3.0	19	8.0	6.5
8	3.5	2.7	20	8.0	7.0
9	4.0	4.0	21	8.9	8.5
10	4.0	3.5	22	9.0	8.0
11	4.5	4.2	23	9.5	8.1
12	4.6	3.5	24	10.0	8.1

我们以(x,y)为坐标把这 24 对数画到平面直角坐标系内(见图 2),称这样的图为散点图. 从图 2 中可以看出,这些点大体上散布在某条直线周围,这说明强度 y 与拉伸倍数 x 之间大致呈线性

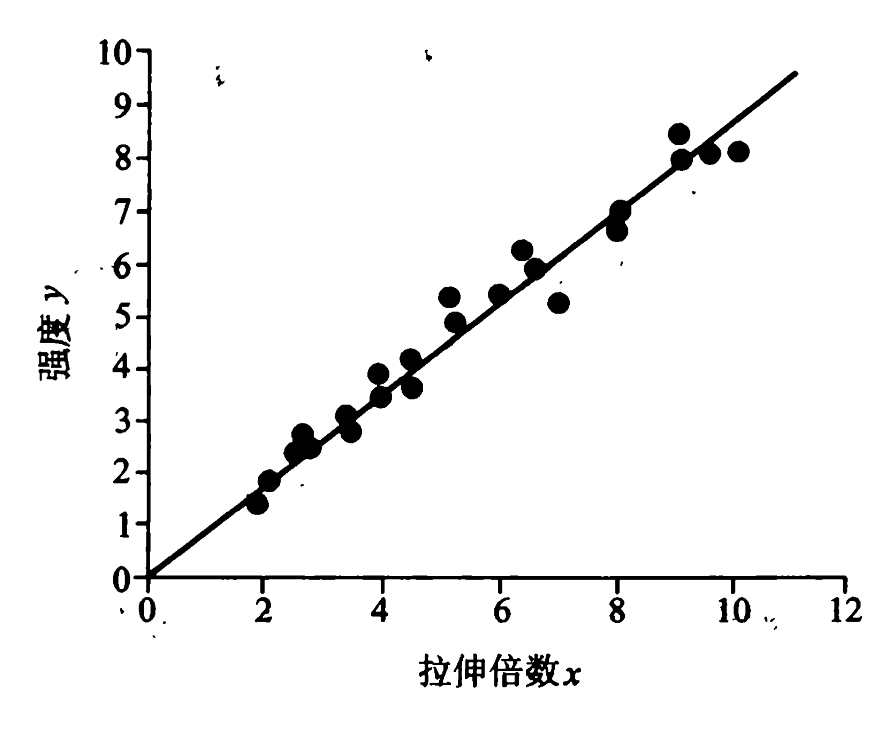


图 2 纤维强度与拉伸倍数

关系：

$$y=a+bx+e \tag{2.1}$$

其中 e 是随机项，a 和 b 是未知的参数．下面的问题就是设法取 a，b 使 e 在某种意义下最小．我们先讨论一般方法，然后再来解决这个例子的问题．

（二）最小二乘法

因为从图中可以看出散点都接近某一条直线，因此就想到应取 a,b 使直线 $y=a+bx$ 最接近图 2 中的 24 个点．怎样衡量一条直线 $y=a+bx$ 与 n 个点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 的接近程度呢？可用

$$Q(a,b)=\sum_{i=1}^{n}\{y_i-(a+bx_i)\}^2 \tag{2.2}$$

的大小来刻画．它是直线与各散点在垂直方向（纵轴方向）上的距离的平方和．$Q(a,b)$ 越小，则直线 $y=a+bx$ 越靠近所述的 n 个点．因此只要取 $\hat{a},\hat{b}$ 使

$$Q(\hat{a},\hat{b})=\min_{a,b} Q(a,b) \tag{2.3}$$

则直线与散点可以很靠近．用这样的方法来确定式(2.1)中的 a，b，通称最小二乘法，意思是使平方和最小的参数估计方法．这样得到的 $\hat{a},\hat{b}$ 称为 a,b 的最小二乘估计．

定理 2.1　对 n 对值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，如果 x_1，$x_2,\cdots,x_n$ 不全相等，则 a,b 的最小二乘估计为

$$\hat{a}=\bar{y}-\hat{b}\,\bar{x}$$

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2.4}$$

其中 $\bar{x}$ 是 $x_1,\cdots,x_n$ 的平均值，$\bar{y}$ 是 $y_1,\cdots,y_n$ 的平均值．

证明：$Q(a,b)=\sum_{i=1}^{n}\{y_i-(a+bx_i)\}^2$

$$= \sum_{i=1}^{n} \{(y_i - \overline{y}) + [\overline{y} - (a + b\overline{x})] - b(x_i - \overline{x})\}^2$$

$$= \sum_{i=1}^{n} (y_i - \overline{y})^2 + n[\overline{y} - (a + b\overline{x})]^2 +$$

$$b^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2b \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})$$

记

$$l_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$l_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$l_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

则

$$\begin{aligned} Q(a,b) &= l_{yy} + n[\overline{y} - (a + b\overline{x})]^2 + b^2 l_{xx} - 2b l_{xy} \\ &= l_{yy} + n[\overline{y} - (a + b\overline{x})]^2 + l_{xx}\left(b - \frac{l_{xy}}{l_{xx}}\right)^2 - \frac{l_{xy}^2}{l_{xx}} \\ &\geqslant l_{yy} - \frac{l_{xy}^2}{l_{xx}} \end{aligned}$$

要使 $Q(a,b)$ 达到最小值必须且只须

$$\overline{y} = a + b\overline{x}$$

$$b = \frac{l_{xy}}{l_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

因此 a,b 的最小二乘估计为(2.4)中的 $\hat{a},\hat{b}$. 证毕.

从定理 2.1 可得 y 与 x 的经验公式是

$$\hat{y} = \hat{a} + \hat{b}x \tag{2.5}$$

其中 $\hat{a},\hat{b}$ 由(2.4)式确定. 在符号 y 上加的帽子记号是为了区别于 y 的实际值. $\hat{y}$ 叫做对应于 x 的回归值,(2.5)式叫做回归方

程，由回归方程决定的直线叫回归直线.

对于例 2.1 的数据，用(2.4)式可以算出

$$\hat{b}=0.859, \hat{a}=0.15$$

因此，强度 y 与拉伸倍数 x 之间的经验公式(回归方程)为

$$\hat{y}=0.15+0.859x$$

这里回归系数 b 估计为 0.859 的意义是，拉伸倍数每增加一个单位，强度平均增加 0.859 个单位.

(三) 线性关系的检验

有了经验公式(2.5)后，是否就可以用它来进行预报和控制了呢？我们必须注意，定理 2.1 除要求各 x_i 不全相等以外对数据并没有作什么假设，因此给了任何一组数据 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 都可以用公式(2.4)来计算最小二乘估计 $\hat{a}, \hat{b}$ 从而得到 y 和 x 之间的一个经验公式，这并不能说明 y 与 x 之间真有近似的线性关系. 如果回归值与真实值之间的差距即拟合误差 $\hat{y}-y$ 确实都比较小，我们就可以认为 y 与 x 之间的确存在线性相关关系. 判断一个随机变量是否比较小只能用假设检验的方法，下面我们就来构造这种检验，称为线性相关关系的检验.

我们先证明一个具有统计意义的平方和分解公式：

引理 2.1 令

$$\hat{y}_i=\hat{a}+\hat{b}x_i \ (i=1,2,\cdots,n)$$

其中 $\hat{a}, \hat{b}$ 为最小二乘估计，则

$$l_{yy}=Q+U \tag{2.6}$$

其中

$$l_{yy}=\sum_{i=1}^{n}(y_i-\overline{y})^2$$

叫 y 的总离差平方和，

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

叫残差平方和，

$$U = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

叫回归平方和.

证明：
$$\begin{aligned}
l_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})]^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \\
&\quad 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})
\end{aligned}$$

其中交叉项

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \\
&= \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)(\hat{a} + \hat{b}x_i - \overline{y}) \\
&= \sum_{i=1}^{n}[y_i - \overline{y} - \hat{b}(x_i - \overline{x})][\hat{b}(x_i - \overline{x})] \\
&= \hat{b}(l_{xy} - \hat{b}l_{xx}) \\
&= \hat{b}l_{xx}\left(\frac{l_{xy}}{l_{xx}} - \hat{b}\right) \\
&= 0
\end{aligned}$$

故(2.6)式成立. 证毕.

(2.6)式左边的 l_{yy} 反映了 $y_1, y_2, \cdots, y_n$ 这 n 个值的分散程度，(2.6)式说明 l_{yy} 可以分成两部分 Q 和 U，U 是拟合值 $\hat{y}_i, i=1, 2, \cdots, n$ 的分散程度，它由回归线上不同的自变量值 x_i 引起，即 U 这一部分平方和可以得到回归式的解释. Q 是拟合残差的平方和，它反映了回归直线与观测数据的接近程度，Q 越小，说明回归的效果越好. 因此在 l_{yy} 的这两部分中，当 Q 所占的比例相当小(相应地，U 所占比例相当大)时，原来的数据中的近似的线性关系才比

较可信. 因此我们用比值 U/Q 的大小来衡量线性相关关系的可信程度,这个比值越大,我们越有把握说两变量之间确实存在线性相关关系.

为了建立统计上的假设检验,需要先作一些数学上的假定. 设数据 (x_i, y_i) 有下列结构

$$y_i = a + bx_i + e_i \quad (i = 1, 2, \cdots, n) \tag{2.7}$$

其中 $e_1, e_2, \cdots, e_n$ 为独立同 $N(0, \sigma^2)$ 分布的随机变量($\sigma^2 > 0$ 未知). 我们来检验假设

$$H_0 : b = 0 \tag{2.8}$$

如果否定了 H_0,就说明 y 与 x 之间有线性相关关系;如不能否定 H_0,说明从现有数据出发不能认为 y 与 x 线性相关. 对 H_0 的检验叫相关性检验. 当 H_0 被否定时,称回归方程(2.5)是显著的. 令

$$F = \frac{U}{Q/(n-2)} \tag{2.9}$$

可以证明(见§5),在 H_0 下 F 服从自由度为 $(1, n-2)$ 的 F 分布. 取否定域

$$W = \{F > \lambda\}$$

其中 λ 为 $F(1, n-2)$ 分布的 $1-\alpha$ 分位数,α 为检验水平. 实际计算 F 时,可用下列公式

$$U = \hat{b} l_{xy} \tag{2.10}$$

$$Q = l_{yy} - U \tag{2.11}$$

其中第二式由(2.6)式可得,第一式是由于

$$\begin{aligned} U &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \\ &= \hat{b}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \hat{b}^2 \cdot l_{xx} \\ &= \hat{b} \cdot l_{xy} \end{aligned}$$

现在回到例 2.1 上来. 用统计量(2.9)进行相关性检验，计算得 $F=436$，查 F 分布表得自由度为(1,22)时 0.95 分位数为 4.30. 而 $F>4.30$，故应否定 $H_0: b=0$，即强度 y 与拉伸倍数 x 之间确实存在线性相关关系，回归方程 $\hat{y}=0.15+0.859x$ 可以用于预测、控制等目的.

(四) 相关系数

在实际工作中，要判断变量 y 与变量 x 之间的线性相关性，人们还常用 y 与 x 的样本相关系数

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

来衡量. r 是把 y 和 x 看成随机变量时其相关系数的矩估计，我们知道两个随机变量 y 和 x 的相关系数 $\rho(x,y)$ 取值在$[-1,1]$内，x,y 独立时 $\rho(x,y)=0$，而 y 与 x 可以互相线性表示时 $|\rho(x,y)|=1$，$|\rho(x,y)|$ 的大小代表了 y 和 x 的线性相关性的强弱. 虽然这里我们把 x 作为非随机的普通变量，我们还是可以用估计量 r 来作为 y 与 x 的线性相关性的度量.

对一元线性回归，我们还可以建立 r 与回归平方和 U、残差平方和 Q 和总离差平方和 l_{yy} 之间的关系：

$$r^2=\frac{U}{l_{yy}}=1-\frac{Q}{l_{yy}}$$

事实上，

$$r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}}=\frac{l_{xy}\cdot\frac{l_{xy}}{l_{xx}}}{l_{yy}}=\frac{l_{xy}\cdot\hat{b}}{l_{yy}}$$

$$=\frac{U}{l_{yy}}=1-\frac{Q}{l_{yy}}$$

所以 r^2 是回归平方和 U 在总离差平方和 l_{yy} 中占的比例，r^2 越大说明 y 与 x 线性相关性越强.

r 与最小二乘估计 $\hat{b}$ 还有如下关系

$$r=\hat{b}\sqrt{\frac{l_{xx}}{l_{yy}}}$$

r 与(2.9)的 F 统计量关系为

$$F=\frac{r^2}{(1-r^2)/(n-2)}$$

故对 $H_0:b=0$ 的检验可通过 r 进行,但临界值仍根据上式用 F 分布来求.

细心的读者可能注意到,在用统计量 F 检验假设 $H_0:b=0$ 时,我们用到了关于模型(2.7)的"假定":随机项 $e_1,e_2,\cdots,e_n$ 相互独立同分布,共同分布是 $N(0,\sigma^2)$(σ 未知). 如何根据数据$(x_i,y_i)(i=1,\cdots,n)$判别这个"假定"是否成立呢? 这又是一个检验问题. 令 $\hat{e}_i=y_i-\hat{y}_i(i=1,\cdots,n)$,这里 $\hat{y}_i=\hat{a}+\hat{b}x_i$($\hat{a},\hat{b}$ 分别是 a,b 的最小二乘估计). 这些 $\hat{e}_i$ 叫作残差. 令

$$\gamma_i=\frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-p_i}}\quad(i=1,\cdots,n),$$

其中

$$\hat{\sigma}=\sqrt{\frac{Q}{n-2}},\quad p_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{l_{xx}}$$

$$Q=\sum_1^n(y_i-\hat{y}_i)^2,\quad l_{xx}=\sum_1^n(x_k-\bar{x})^2.$$

数学上可以证明[15],如果上述"假定"成立,则 $\gamma_1,\cdots,\gamma_n$ 近似地相互独立,且各 γ_i 近似地服从标准正态分布. 从而在 n 比较大时,$\gamma_1,\cdots,\gamma_n$ 中大约有$[0.95n]$个 γ_i 满足$|\gamma_i|\leqslant 2$

反之,若根据数据计算出 $\gamma_1,\cdots,\gamma_n$ 后发现满足不等式$|\gamma_i|\leqslant 2$ 的 γ_i 的个数与$[0.95n]$差别很大,则表明上述的"假定"不成立,即不能认为模型(2.7)的随机项 $e_1,e_2,\cdots,e_n$ 是相互独立同分布的且共同分布是 $N(0,\sigma^2)$. 这时用统计量 F 检验 $H_0:b=0$ 就缺乏根据了.

只要上述的"假定"未被否定,我们就认为"假定"成立. 这方面的深入研究属残差分析的范围,参看[15].

(五) 预测

有了回归直线,当自变量 $x=x_0$ 时,相应的 y_0 是多少? 自然

用 $\hat{y}_0=\hat{a}+\hat{b}x_0$ 去预测 y_0（也可说用 $\hat{y}_0$ 作为 y_0 的估计）. 这里要注意的是，由于 y_0 是随机变量，故 $\hat{y}_0$ 不是第二章中对参数的估计，而是对随机变量的估计. 我们也可以构造 y_0 的置信区间. 可以证明，在一定条件下（见第 5 节）.

$$T=\frac{y_0-\hat{y}_0}{\sqrt{dQ/(n-2)}}$$

服从 $n-2$ 个自由度的 t 分布，其中

$$d=1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}},\bar{x}=\frac{1}{n}\sum_1^n x_i$$

给定 $\alpha\in(0,1)$. 从 t 分布表中可找到临界值 λ 满足

$$P(|T|\leqslant\lambda)=1-\alpha$$

于是可得 y_0 的 $1-\alpha$ 水平置信区间

$$\left[\hat{y}_0-\lambda\sqrt{dQ/(n-2)},\hat{y}_0+\lambda\sqrt{dQ/(n-2)}\right] \tag{2.12}$$

当自变量 x_0 在横轴上变化时，置信区间的上、下限的相应变化形成双曲线（如图 3 所示）.

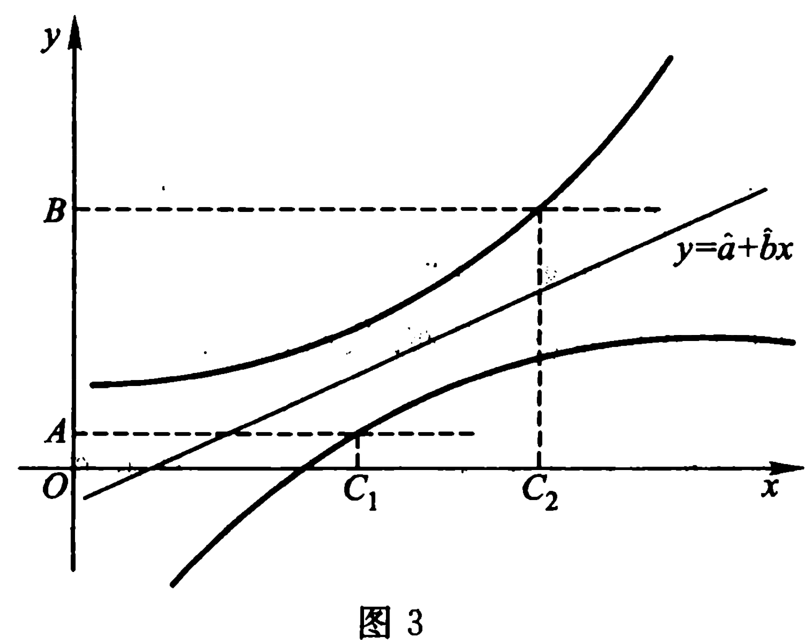


图 3

例如，$\alpha=0.05$ 时，针对表 2.1 的数据，$n=24,\lambda=2.074,l_{xx}=152.3$

$$d=\frac{25}{24}+\frac{(x_0-5.313)^2}{152.3},\quad Q=5.66,$$

于是 y_0 的置信区间是

$$\left[\hat{y}_0 - 1.052\sqrt{\frac{25}{24} + \frac{(x_0 - 5.313)^2}{152.3}}, \hat{y}_0 + 1.052\sqrt{\frac{25}{24} + \frac{(x_0 - 5.313)^2}{152.3}}\right]$$

(置信水平是 0.95).

(六)控制

如果想使 y_0 落在给定的范围$[A,B]$内,x_0 应该取什么值?这就是控制问题. 从(2.12)知道,若要以 $1-\alpha$ 的概率保证 y_0 取值在$[A,B]$内,则 x_0 必须且只需满足不等式

$$\hat{a}+\hat{b}x_0-\lambda\sqrt{dQ/(n-2)}\geqslant A$$

$$\hat{a}+\hat{b}x_0+\lambda\sqrt{dQ/(n-2)}\leqslant B$$

解上述不等式(注意 d 也是 x_0 的函数!)可以得到 $x_0\in[C_1,C_2]$. 从图 3 看出,C_1,C_2 分别是 $y=A$,$y=B$ 两条直线与双曲线的交点的横坐标. 在有些情况下,如果$[A,B]$过窄,则满足上述不等式的 x_0 不存在.

(七)一元齐次线性回归

在某些实际问题里,根椐事物的实际情况,在研究线性相关性时可以肯定"常数项"是 0. 即应考虑下列模型:

$$Y=bx+e \tag{2.13}$$

其中 e 是随机项,$Ee=0$.

设有数据$(x_1,y_1),\cdots,(x_n,y_n)$,如何估计系数 b? 和前面一样,应找 b 使平方和 $\sum_1^n(y_i-bx_i)^2$ 达到最小值. 易知当且仅当 $b=\hat{b}$ 时这个平方和达到最小值,这里

$$\hat{b}=\frac{\sum_1^n x_iy_i}{\sum_1^n x_i^2} \tag{2.14}$$

这个 $\hat{b}$ 仍叫作 b 的最小二乘估计. 于是有回归方程

$$y=\hat{b}x \tag{2.15}$$

这个方程是否有意义呢? 还应检验假设 $H_0:b=0$. 只当这个假设

被拒绝时方程(2.15)才有意义，才可用来进行预测或控制.

怎样检验 $H_0: b=0$ 呢？和以前一样，应假定数据有下列结构：

$$y_i = bx_i + e_i \quad (i=1,\cdots,n),$$

这里 $e_1, e_2, \cdots, e_n$ 相互独立同分布，共同分布是 $N(0,\sigma^2)$(σ 未知).

显然，在 H_0 下 $y_1, y_2, \cdots, y_n$ 独立同分布，共同分布是 $N(0, \sigma^2)$. 于是 $\overline{y}$ 与 $\sum_1^n (y_i - \overline{y})^2$ 相互独立，$\sqrt{n}\frac{\overline{y}}{\sigma} \sim N(0,1)$，$\frac{1}{\sigma^2}\sum_1^n (y_i - \overline{y})^2 \sim \chi^2(n-1)$. 从而统计量

$$F = \frac{n(\overline{y})^2}{\sum_1^n (y_i - \overline{y})^2/(n-1)} \tag{2.16}$$

服从第 1 自由度是 1，第 2 自由度是 $n-1$ 的 F 分布. 因而在 H_0 下查 F 分布表有临界值 λ 满足 $P(F>\lambda)=\alpha$. 所以在 F 的值大于 λ 时拒绝 H_0. 此时所得到的回归方程才是有意义的.

例 2.2 哈勃定律是 20 世纪最惊人的科学发现之一，它说明宇宙正在膨胀，各星系在以巨大的速度彼此分离. 设 v 是一个星系相对于另一任意星系的速度，d 为此二星系之间的距离，则哈勃定律为

$$v = Hd$$

其中 H 称为哈勃常数.

表 2.2 列出了 11 个星团距太阳系的距离和相对速度，我们由此来估计哈勃常数. 易见 H 的最小二乘估计，即使 $\sum_{i=1}^n (v_i - Hd_i)^2$ 达最小的 H 值为

$$\hat{H} = \frac{\sum_{i=1}^{11} d_i v_i}{\sum_{i=1}^{11} d_i^2} = 0.035\,44$$

图 4 为 11 个点及回归线 $v=0.035\,44d$.

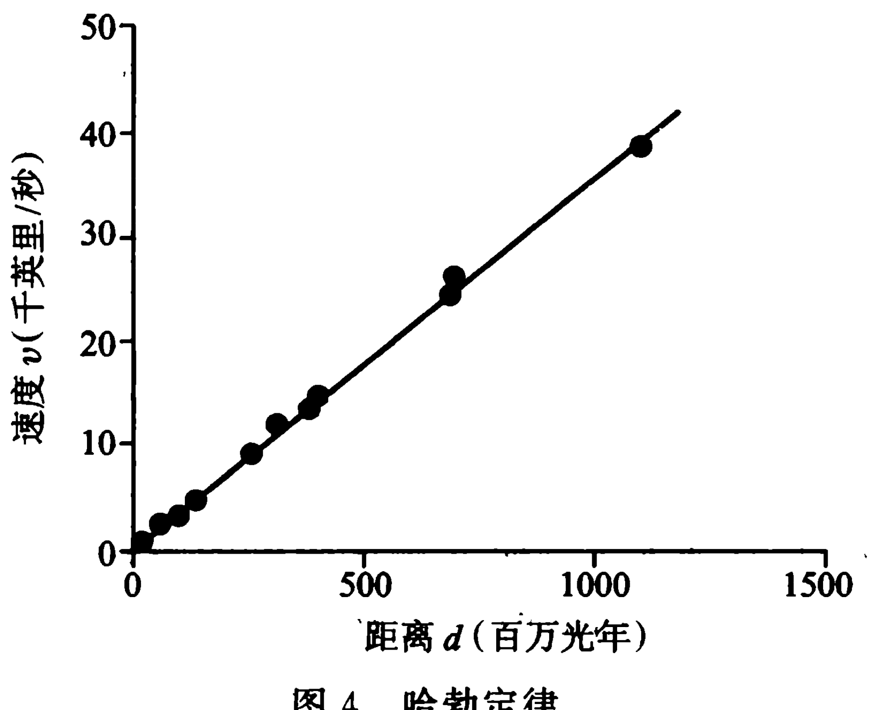


图 4　哈勃定律

利用(2.16)可算出统计量 $F=14.69$,查 F 分布表知 $F(1,10)$ 的 0.95 分位数是 4.96 现在 $F>4.96$. 故所得到的回归线是有意义的.

表 2.2　星团距离、速度数据

星　团　名	距离(百万光年)	速度(千英里①/秒)
室女座	22	0.75
飞马座	68	2.4
英仙座	108	3.2
后发座	137	4.7
大熊座一号	255	9.3
狮子座	315	12.0
北冕座	390	13.4
双子座	405	14.4
牧夫座	685	24.5
大熊座二号	700	26.0
长蛇座	1100	38.0

① 英里为非法定计量单位,1 英里≈1 609.344 米.

哈勃常数对天文学家十分重要. 根据近代的宇宙大爆炸理论,哈勃定律是大爆炸的结果,而哈勃常数的倒数则是宇宙的年龄. 注意上面估出的 H 的单位是英里/光年·秒,换算为秒$^{-1}$后可由$\frac{1}{H}$知道宇宙的年龄. 表 2.2 是 20 年代哈勃所用的数据,由此得到的宇宙年龄是 52 亿年左右. 由于现在测量星系间距离的技术有了很大改进,所以用现在的数据算出的宇宙年龄约为 150 亿年.

例 2.3 哺乳动物出生后开始走动的时间 x 与开始玩耍的时间 y 是有关的,有许多学者研究过这两者的关系. 表 2.3 列出了 11 种哺乳动物开始走动的时间与开始玩耍的时间的数据. 画出散点图后,可以看出数据点好像落在类似 $y=ax^b$ 的幂曲线周围.

令 $y'_i=\ln y_i, x'_i=\ln x_i$,对 y'_i, x'_i 进行线性回归拟合得回归方程

$$\hat{y}'=1.689+0.561x' \tag{2.17}$$

相当于

$$\hat{y}=5.42x^{0.561} \tag{2.18}$$

(2.17)式对应的 F 统计量为 27.438,在水平 0.000 5 下方程(2.17)是显著的.

表 2.3 哺乳动物数据

种　别	开始走动时间 x(天数)	开始玩耍时间 y(天数)
人	360	90
大猩猩	165	105
猫	21	21
家犬	23	26
挪威鼠	11	14
乌鸦	18	28
混血猕猴	18	21
黑猩猩	150	105

续表

种　别	开始走动时间 x(天数)	开始玩耍时间 y(天数)
松鼠猴	45	68
花鼠	45	75
白脸猴	18	46

如果直接对 y_i 和 x_i 拟合最小二乘估计得回归直线

$$\hat{y}=35.81+0.2347x \tag{2.19}$$

它对应的 F 统计量只有 9.506，在水平 0.01 下方程(2.19)不显著，在水平 0.05 下显著，说明 y 和 x 的关系用幂函数表示比用线性关系要好. 从图 5 也可以看到这一点.

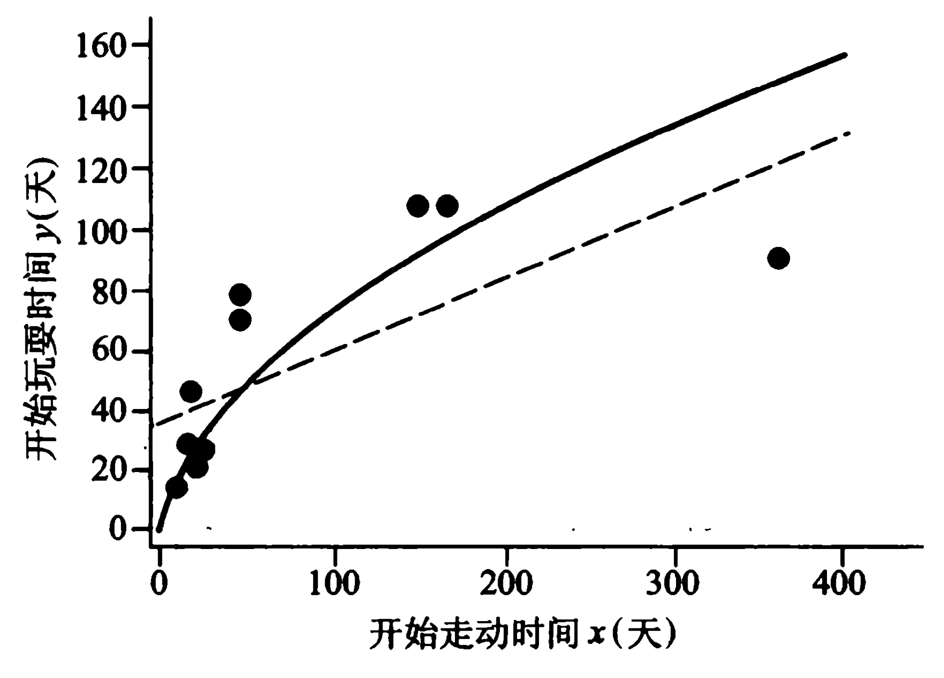


图 5　哺乳动物开始走动时间与开始玩耍时间

上一节里我们已经说过，线性回归之所以重要，不仅是因为它数学上容易处理，实际中经常出现，还因为实际中许多数据经过适当变换后可以化为用线性回归的形式. 变换的选择首先应靠有关专业知识，如果有关专业知识不能提供这方面的信息则可由数据散点图的形状来试选适当的函数对应的变换.

一元多项式 $y=a_0+a_1x+\cdots+a_px^p$ 总可化为多元线性回归

形式.

下面图 6 至图 12 是其他一些常用的函数曲线.

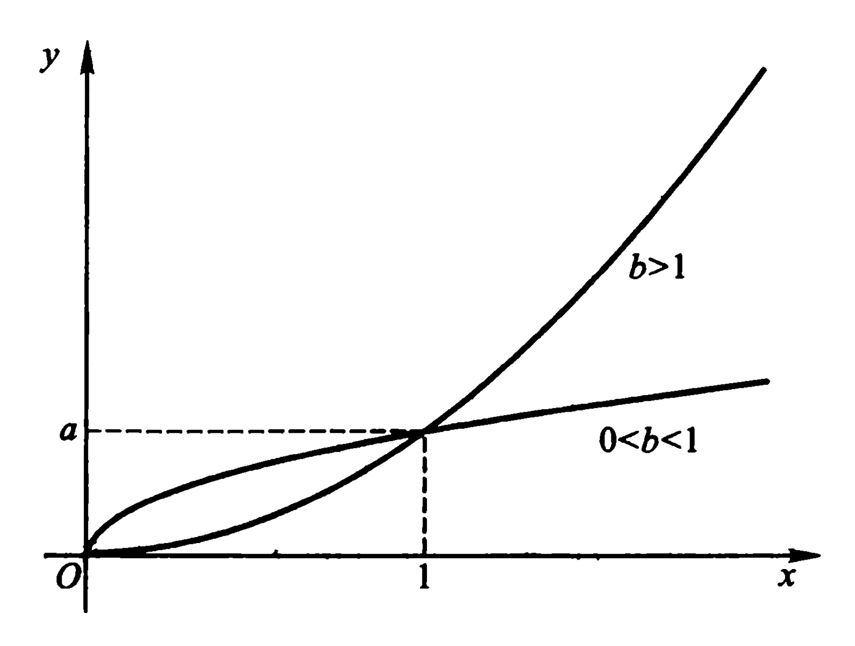


图 6　幂函数 $y=ax^{b}(b>0)$

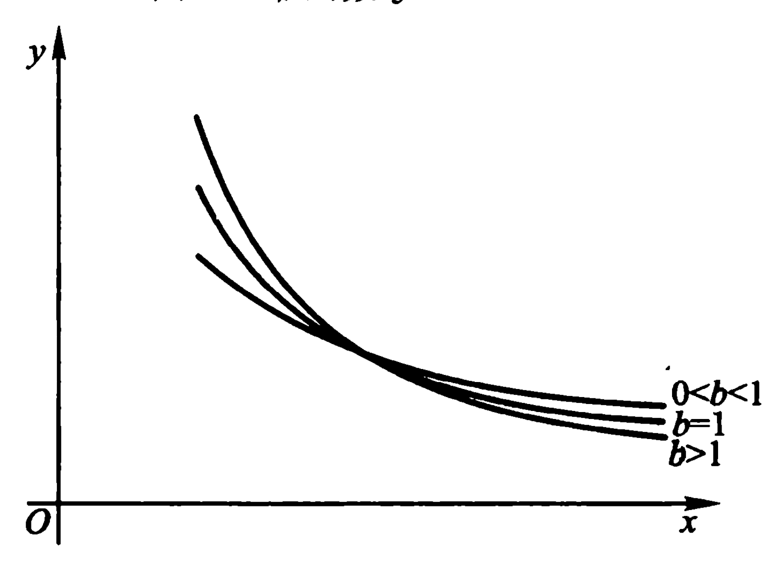


图 7　幂函数 $y=ax^{-b}(b>0)$

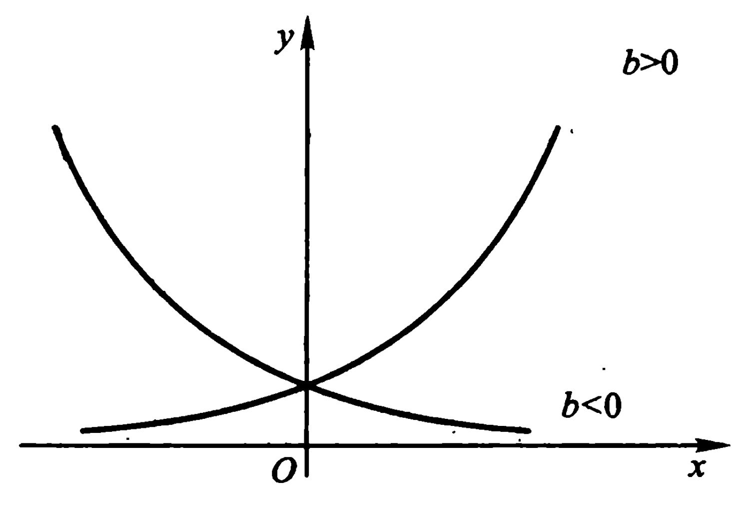


图 8　指数函数 $y=ae^{bx}$

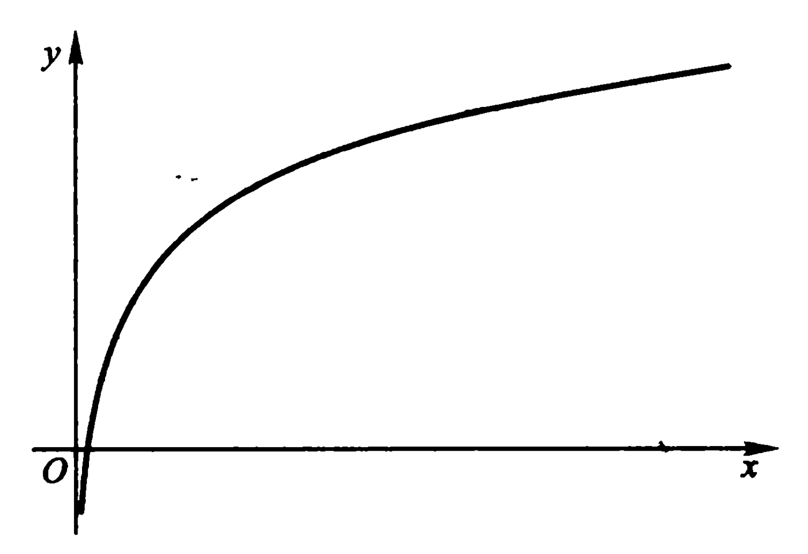


图 9　对数函数 $y=a+b\ln x(b>0)$

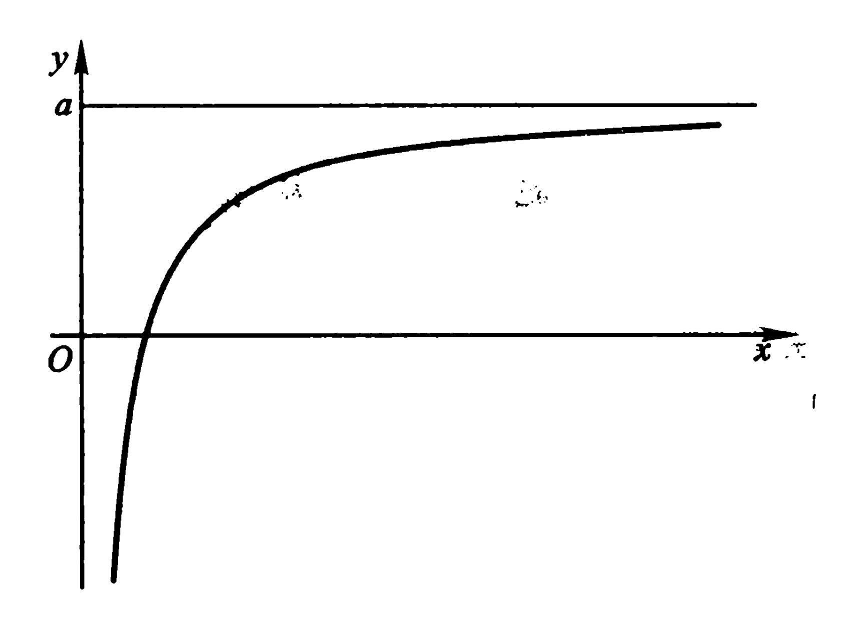


图 10　函数 $y=a-b/x(b>0)$

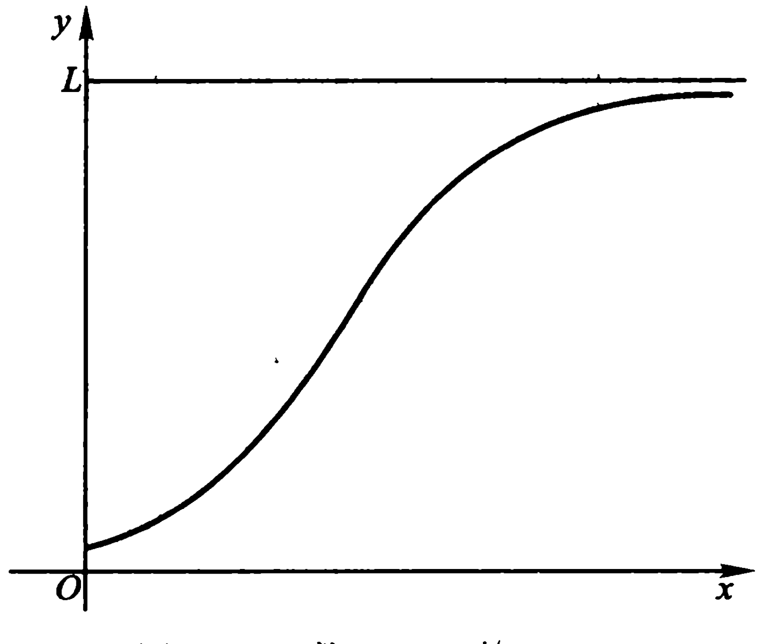


图 11　函数 $y=ae^{-b/x}(b>0)$

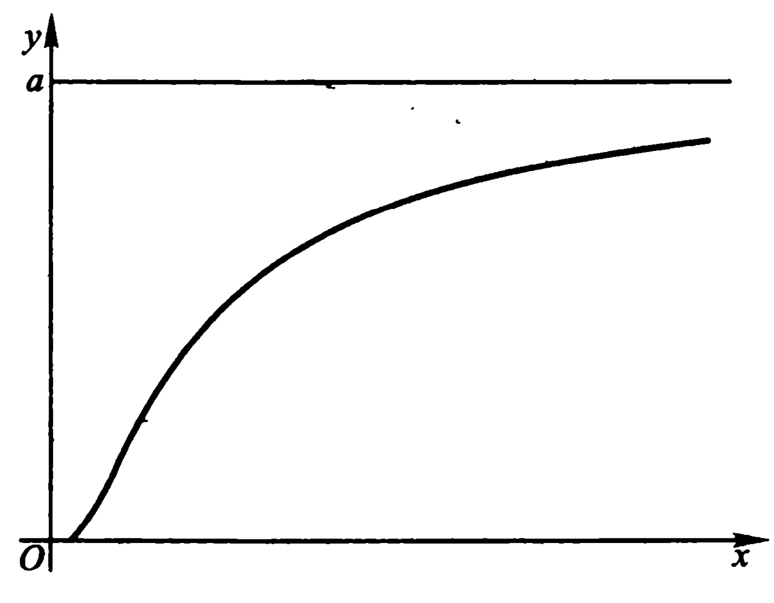


图 12　逻辑曲线 $y=\frac{L}{1+e^{a-bx}}(b>0)$

§3　线性模型的参数估计

(一) 线性模型

多元线性回归的一般形式是

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+e$$

不失一般性，总可以设 $\beta_0=0$，因为可以形式地引入自变量 $x_0\equiv 1$，从而 $y=\sum_{i=0}^{p}\beta_i x_i+e$. 所以在理论研究时可以假定

$$y=\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+e \tag{3.1}$$

多元线性回归的出发点，正如§1中所述，是表1.1那样的数据，即

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{matrix} \tag{3.2}$$

其中 y_i 是 $x_1=x_{i1}, x_2=x_{i2},\cdots,x_p=x_{ip}$ 时 y 的观测值($i=1,2,\cdots,n$)，$n>p$.

我们假设上述数据适合下列模型，称为线性模型：

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\
&\cdots\cdots\cdots\cdots \\
y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n
\end{aligned}
\tag{3.3}
$$

其中 $e_1, e_2, \cdots, e_n$ 是随机变量(这样(3.3)式中的 $y_1, y_2, \cdots, y_n$ 也看成随机变量,而(3.2)中的 $y_1, y_2, \cdots, y_n$ 为它们的观测值,在不引起误会的情况下,在记号上不区分随机变量及其观测值),x_{ij}, $j=1,2,\cdots,p, i=1,2,\cdots,n$ 是已知的实数(即把 $x_1, x_2, \cdots, x_p$ 看作非随机的普通变量),$\beta_1, \beta_2, \cdots, \beta_p$ 是未知参数.

对随机项 $e_1, e_2, \cdots, e_n$ 常作下列两种假定:

假定 A: $\quad Ee_i = 0 \quad (i=1,2,\cdots,n)$

$Ee_i e_j = 0 \quad (i \neq j)$

$Ee_i^2 = \sigma^2 \quad (i=1,2,\cdots,n)$

其中 σ 为未知正数.

假定 B: $e_1, e_2, \cdots, e_n$ 相互独立同分布且 $e_1 \sim N(0, \sigma^2)$,其中 σ 是未知正数.

显然假定 B 比假定 A 强,线性模型研究只要不涉及假设检验和置信区间一般只用到假定 A.

为了表达简捷并便于使用线性代数的知识,我们大量使用矩阵和向量记号. 记

$$
Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = (x_{ij})_{n \times p}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}
$$

$$
\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}
$$

则模型(3.3)可改写为

$$Y = X\beta + e \tag{3.4}$$

其中 e 为 n 维随机向量，X 为 $n\times p$ 常数阵，β 为 p 维未知参数向量．假定 A 可表达为

$$Ee = 0, \mathrm{Cov}(Y, Y) = \sigma^2 I$$

其中 I 为单位阵；$\mathrm{Cov}(\xi, \eta)$ 表示 ξ 与 η 的协方差阵．假定 B 可表示为

$$e \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 I\right)$$

或

$$Y \sim N(X\beta, \sigma^2 I)$$

以下我们还用到一些线性代数中的记号．设 a 为 R^k 中的向量，M 为 R^k 的子空间，则记 $\|a\|$ 为 a 的长度，记 $\underset{M}{\mathrm{Proj}}\, a$ 为向量 a 在子空间 M 中的投影向量．设 A 为一 $m\times k$ 矩阵，则 $\mu(A)$ 表示由 A 的 k 个列张成的 $\mathbf{R}^m$ 的子空间．向量 a 可以有正交分解

$$a = \underset{M}{\mathrm{Proj}}\, a + (a - \underset{M}{\mathrm{Proj}}\, a)$$

其中 $(a - \underset{M}{\mathrm{Proj}}\, a) \perp M$.

（二）线性模型参数的最小二乘估计

下面研究线性模型(3.4)中参数 β 的估计问题．有了一般理论，不仅可以用于回归分析，还可以用于方差分析、试验设计、数据平滑等多种统计问题．

已知矩阵 X 及观测向量 Y，如何估计(3.4)中的 β? 应该选 β 使得(3.4)中的 e"很小"．受到上一节一元回归的启发，一种自然的想法是使

$$\begin{aligned} Q(\beta) &= \sum_{i=1}^{n}\{y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\}^2 \\ &= \|Y - X\beta\|^2 \end{aligned} \tag{3.5}$$

达到最小.

定义 3.1 对线性模型(3.4),称 $\hat{\beta}$ 是 β 的最小二乘估计,如果对一切 β,有

$$Q(\beta) \geqslant Q(\hat{\beta})$$

下面的定理论述了最小二乘估计的存在性及其求法.

定理 3.1 (1) 最小二乘估计一定存在.

(2) $\hat{\beta}$ 是最小二乘估计的充要条件是 $\hat{\beta}$ 适合方程(称为模型(3.4)的正规方程)

$$X'X\beta = X'Y \tag{3.6}$$

证明:(1) 注意到 X 的各列张成的线性空间 $\mu(X)$ 可表示为

$$\mu(X) = \{\eta \in R^n : \eta = X\beta, \beta \in R^p\}$$

令

$$\xi = \operatorname*{Proj}_{\mu(X)} Y$$

则由投影变换的性质知

$$\|Y - \xi\| \leqslant \|Y - X\beta\|, \quad \forall \beta \in R^p$$

而 $\xi \in \mu(X)$,从而必存在 $\hat{\beta} \in R^p$ 使 $X\hat{\beta} = \xi$,即

$$\|Y - X\hat{\beta}\| \leqslant \|Y - X\beta\|, \quad \forall \beta \in R^p$$

即

$$Q(\hat{\beta}) \leqslant Q(\beta), \quad \forall \beta \in R^p$$

因此 $\hat{\beta}$ 就是 β 的最小二乘估计.

(2) 仍记 $\xi = \operatorname*{Proj}_{\mu(X)} Y$,则若 $X\hat{\beta} = \xi$,当然有

$$Q(\hat{\beta}) \leqslant Q(\beta), \quad \forall \beta \in R^p \tag{3.7}$$

反过来,若 $\hat{\beta}$ 使(3.7)成立,注意到

$$Y = \xi + (Y - \xi)$$

$$Y - X\hat{\beta} = (\xi - X\hat{\beta}) + (Y - \xi)$$

$$\|Y - X\hat{\beta}\|^2 = \|Y - \xi\|^2 + \|\xi - X\hat{\beta}\|^2$$

若 $X\hat{\beta} \neq \xi$ 则 $Q(\hat{\beta}) > \|Y - \xi\|^2$,而 $\xi \in \mu(X)$,与(3.7)矛盾,故必有 $X\hat{\beta} = \xi$. 因此(3.7)成立的充要条件是

$$X\hat{\beta}=\xi=\underset{\mu(X)}{\text{Proj}}\, Y$$

即 $Y-X\hat{\beta}\perp\mu(X)$，即 $Y-X\hat{\beta}$ 垂直于 X 的每一列，这可表为

$$X'(Y-X\hat{\beta})=0$$

即(3.6)式成立　　　　　　　　　　　　　　　　　　证毕.

注　我们可以不用线性空间的投影理论，直接利用矩阵的基本运算给出本定理的初等证明.

设 $\tilde{\beta}=(\tilde{\beta}_1,\cdots,\tilde{\beta}_p)'$，从(3.5)知

$$\begin{aligned}Q(\beta)&=(Y-X\beta)'(Y-X\beta)=(Y-X\tilde{\beta}+X(\tilde{\beta}-\beta))'(Y-X\tilde{\beta}+X(\tilde{\beta}-\beta))\\&=(Y-X\tilde{\beta})'(Y-X\tilde{\beta})+(X(\tilde{\beta}-\beta))'X(\tilde{\beta}-\beta)+2(X(\tilde{\beta}-\beta))'(Y-X\tilde{\beta})\\&=Q(\tilde{\beta})+\|X(\tilde{\beta}-\beta)\|^2-2(\tilde{\beta}-\beta)'(X'X\tilde{\beta}-X'Y).\end{aligned}\tag{3.8}$$

由此可见，若 $\tilde{\beta}$ 满足方程(3.6)，则 $Q(\beta)\geqslant Q(\tilde{\beta})$，即 $\tilde{\beta}$ 是 β 的最小二乘估计.

我们指出方程(3.6)一定有解. 实际上，线性方程组(3.6)的系数矩阵 $X'X$ 与 X 有相同的秩，而增广矩阵 $(X'X,X'Y)=X'(X,Y)$ 的秩不超过 X 的秩，所以增广矩阵与系数矩阵有相等的秩. 依线性方程的解的存在定理，知方程(3.6)一定有解. 这就证明了最小二乘估计一定存在. 设 $\tilde{\beta}$ 是方程(3.6)的任何一个解，$\hat{\beta}$ 是 β 的任何最小二乘估计，从(3.8)知 $Q(\hat{\beta})=Q(\tilde{\beta})+\|X(\tilde{\beta}-\hat{\beta})\|^2$ 由于 $Q(\hat{\beta})\leqslant Q(\tilde{\beta})$，故 $X(\tilde{\beta}-\hat{\beta})=O$，即 $X\tilde{\beta}=X\hat{\beta}$，从而 $X'X\hat{\beta}=X'X\tilde{\beta}=X'Y$. 这表明 $\hat{\beta}$ 一定满足方程(3.6). 这就完成了定理 3.1 的初等证明.

顺便指出，本章定理的证明中凡是用到线性空间的投影理论的都可利用矩阵知识避免比较抽象的投影理论. 当然，“投影理论”也有其优点：便于从几何上把握研究对象，表达简洁，且便于推广到无穷维空间(时间序列的研究就涉及无穷维线性空间).

以下恒用 $\hat{\beta}$ 表示 β 的最小二乘估计. 易见

$$Q(\hat{\beta})=Y'Y-Y'X\hat{\beta}\tag{3.9}$$

事实上，因 $Y=X\hat{\beta}+(Y-X\hat{\beta})$，所以

$$\begin{aligned}Q(\hat{\beta})&=\|Y-X\hat{\beta}\|^2=\|Y\|^2-\|X\hat{\beta}\|^2\\&=Y'Y-\hat{\beta}'X'X\hat{\beta}\\&=Y'Y-Y'X\hat{\beta}\end{aligned}$$

最小二乘估计可能不唯一. 不过下面的推论中的条件在大多数情况下是满足的.

系 3.1 若 X 满秩(即 X 的秩等于 p),则最小二乘估计唯一且

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{3.10}$$

$$Q(\hat{\beta})=Y'(I-X(X'X)^{-1}X')Y \tag{3.11}$$

这是因为 X 满秩则 $X'X$ 满秩(实际是正定的),因此正规方程有唯一解(3.10). 式(3.11)只要把(3.10)代入(3.9)即可导出.

注意现在我们还没有用到假定 A 或假定 B,我们只知道一定可以求出最小二乘估计,但不知其优良性如何. 下面的定理为了讨论估计的性质就使用了对随机项 e 的假定 A.

定理 3.2 对线性模型(3.4),设 X 满秩,且假定 A 成立,则对 β 的最小二乘估计 $\hat{\beta}$ 有下列结论:

(1) $E\hat{\beta}=\beta$

(2) $\operatorname{Cov}(\hat{\beta},\hat{\beta})=\sigma^2(X'X)^{-1}$

(3) $EQ(\hat{\beta})=(n-p)\sigma^2$

证明:
$$\begin{aligned}E\hat{\beta}&=E[(X'X)^{-1}X'Y]=(X'X)^{-1}X'EY\\&=(X'X)^{-1}X'X\beta=\beta\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(\hat{\beta},\hat{\beta})&=\operatorname{Cov}((X'X)^{-1}X'(X\beta+e),(X'X)^{-1}X'(X\beta+e))\\&=\operatorname{Cov}((X'X)^{-1}X'e,(X'X)^{-1}X'e)\\&=(X'X)^{-1}X'\operatorname{Cov}(e,e)X(X'X)^{-1}\\&=\sigma^2(X'X)^{-1}\end{aligned}$$

记 $A=I-X(X'X)^{-1}X'$,由(3.10)知

$$\begin{aligned}EQ(\hat{\beta})&=E(Y'AY)=E\operatorname{tr}(Y'AY)\\&=E\operatorname{tr}(AYY')\\&=\operatorname{tr}(A\cdot E(YY'))\end{aligned}$$

而

$$\begin{aligned}E(YY')&=\operatorname{Cov}(Y,Y)+X\beta\beta'X'\\&=\sigma^2I+X\beta\beta'X'\end{aligned}$$

又

$$AX=(I-X(X'X)^{-1}X')X=0$$

所以

$$\begin{aligned}EQ(\hat{\beta})&=\sigma^2\operatorname{tr}(A)\\&=\sigma^2\{\operatorname{tr}(I_n)-\operatorname{tr}[X(X'X)^{-1}X']\}\\&=\sigma^2\{n-\operatorname{tr}[(X'X)^{-1}X'X]\}\\&=\sigma^2(n-p)\end{aligned}$$

证毕.

定理 3.2 说明在假定 A 成立及 X 满秩的条件下，最小二乘估计 $\hat{\beta}$ 是参数 β 的无偏估计，而

$$\hat{\sigma}^2 \overset{\text{d}}{=} \frac{1}{n-p}Q(\hat{\beta})$$

是 σ^2 的无偏估计.

若 X 不满秩，则 β 未必有无偏估计，请看下例.

例 3.1 设有两个物体分别重 β_1、β_2，β_1、β_2 未知，把它们一起放在天平的一端称 n 次，第 i 次结果为 y_i. 这时任何一个稍具常识的人都知道这样称不出两个物体各自的重量. 从数学观点来看，我们设

$$y_i=\beta_1+\beta_2+e_i\quad(i=1,2,\cdots,n)$$

其中 $e_1,e_2,\cdots,e_n$ 独立同 $N(0,\sigma^2)$ 分布，σ 未知. 这里矩阵 X 为 $n\times 2$阵，两列的元素都是 1，故不满秩. 凭直观我们知道 β_1 和 β_2 肯定没有无偏估计，对此还可严格证明. 用反证法，如果 β_1 有无偏估计 $\varphi(y_1,y_2,\cdots,y_n)$，即

$$E\varphi(y_1,y_2,\cdots,y_n)=\beta_1$$

而

$$\begin{aligned}&E\varphi(y_1,y_2,\cdots,y_n)=E\varphi(\beta_1+\beta_2+e_1,\cdots,\beta_1+\beta_2+e_n)\\&=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\varphi(u_1,\cdots,u_n)(2\pi\sigma^2)^{-\frac{n}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(u_i-\beta_1-\beta_2)^2}\mathrm{d}u_1\cdots\mathrm{d}u_n\\&=g(\beta_1+\beta_2)=\beta_1\quad(\forall\,\beta_1,\beta_2)\end{aligned}$$

这是不可能的，因 $g(0)=g(0+0)=0=g(1-1)=1$ 不可能成立.

所以 β_1 不存在无偏估计，同理 β_2 也没有无偏估计.

例 3.2 表 3.1 是某个化学试验中，每隔 5 分钟对温度进行一次测量的数据，我们要研究温度随时间变化的规律. 为了简化计算，定义

$$x_i=\frac{t_i-30}{5}$$

$$y_i=(T_i-99.0)\cdot 10$$

容易看出，数据 x_i 与 y_i 的关系不是线性的，我们来作二次曲线拟合. 设

$$y_i=a+bx_i+cx_i^2+e_i$$

表 3.1 试验温度数据

时间 t(分)	x	温度 T(℃)	y	x^2	x^4
5	−5	99.2	2	25	625
10	−4	99.7	7	16	256
15	−3	99.9	9	9	81
20	−2	100.2	12	4	16
25	−1	100.3	13	1	1
30	0	100.4	14	0	0
35	1	100.4	14	1	1
40	2	100.3	13	4	16
45	3	100.0	10	9	81
50	4	99.8	8	16	256
55	5	99.4	4	25	625

由正规方程的一般形式 $X'X\beta=X'Y$ 可得上式的 a，b，c 的最小二乘估计的正规方程为

$$\sum y_i=na+b\sum x_i+c\sum x_i^2$$

$$\sum x_iy_i=a\sum x_i+b\sum x_i^2+c\sum x_i^3$$

$$\sum x_i^2y_i=a\sum x_i^2+b\sum x_i^3+c\sum x_i^4$$

经过计算得到

$$106=11a+110c$$
$$20=110b$$
$$688=110a+1\,958c$$

解出 $\hat{a},\hat{b},\hat{c}$ 得 y 对 x 的二次曲线回归式

$$\hat{y}=13.972\,1+0.181\,8x-0.433\,6x^2$$

可算出残差平方和

$$\begin{aligned} Q &= \|Y-X\hat{\beta}\|^2=\sum y_i^2-\hat{a}\sum y_i-\hat{b}\sum x_iy_i-\hat{c}\sum x_i^2y_i \\ &=1\,188-13.972\,1\times 106-0.181\,8\times 20+0.433\,6\times 688 \\ &=1.638\,2 \end{aligned}$$

随机项方差估计

$$\hat{\sigma}^2=Q/(11-3)=0.204\,8$$

又可算出

$$(X'X)^{-1}=\begin{pmatrix} 0.207\,5 & 0 & -0.011\,66 \\ 0 & 0.009\,091 & 0 \\ -0.011\,66 & 0 & 0.001\,166 \end{pmatrix}$$

所以 $\mathrm{Var}\,\hat{a},\mathrm{Var}\,\hat{b},\mathrm{Var}\,\hat{c}$ 的估计分别为 0.004 25,0.001 862,0.000 238 8.

把变量 x 和 y 用原来的时间 t 和温度 T 代替,得经验公式

$$\hat{T}=98.727+0.107\,7t-0.001\,734t^2$$

为了找出最高温度出现的时间,解方程

$$\frac{\mathrm{d}\hat{T}}{\mathrm{d}t}=0.107\,7-0.003\,468t=0$$

得 $t=31.06$(分)时温度最高.

(三) 线性可估性

单个 β_i 可能没有无偏估计,现在研究 β 的线性函数 $g(\beta)=c'\beta$ 何时有无偏估计,这里

$$c'=(c_1c_2\cdots c_p)$$

为数值向量.

定义 3.2 称 $c'\beta$ 是(线性)可估的,若存在 Y 的线性函数 $a'Y=\sum_{i=1}^{n}a_iy_i$ 使得

$$Ea'Y=c'\beta \quad (\forall \beta \in R^p)$$

如果 X 满秩，则 $\forall c \in R^p$ 有

$$Ec'\hat{\beta}=E\{[c'(X'X)^{-1}X']Y\}=c'\beta$$

即这时 $c'\beta$ 必可估，$a'=c'(X'X)^{-1}X'$. 下面的定理讨论了一般情况下的可估性.

定理 3.3 对线性模型(3.4)，设假定 A 成立，为了 $c'\beta$ 可估，必须且只须 c' 是 X 的行的线性组合，即 $c \in \mu(X')$.

证明：充分性. 设 $c'=a'X(a \in R^n)$，则 $c'\beta=a'X\beta$，从而

$$Ea'Y=a'EY=a'X\beta=c'\beta \quad (\forall \beta \in R^p)$$

因此 $c'\beta$ 可估.

必要性. 设 $c'\beta$ 可估，则存在 a' 使

$$Ea'Y=c'\beta \quad (\forall \beta \in R^p)$$

即

$$a'X\beta=c'\beta, \quad \forall \beta \in R^p$$

$$(a'X-c')\beta=0, \quad \forall \beta \in R^p$$

从而

$$a'X=c'$$

即 c' 是 X 的行的线性组合. 证毕.

如果 $c'\beta$ 可估，则它的无偏估计可以通过 β 的最小二乘估计(一定存在)来构造.

* **定理 3.4**(高斯-马尔科夫) 对线性模型(3.4)，设假定 A 成立，$\hat{\beta}$ 是 β 的最小二乘估计，那么如果 $c'\beta$ 线性可估，$c'\hat{\beta}$ 必为 $c'\beta$ 的唯一的最小方差线性无偏估计.

证明：设 $a'Y$ 是 $c'\beta$ 的一个无偏估计. 令

$$a^*=\operatorname*{Proj}_{\mu(X)} a, \quad \tilde{a}=a-a^*$$

则 $\tilde{a} \perp \mu(X)$，即

$$\tilde{a}'X=0$$

由

$$a'Y=(a^*)'Y+\tilde{a}'Y$$

两边取期望得

$$\begin{aligned} c'\beta &= E[(a^*)'Y] + \tilde{a}'X\beta \\ &= E[(a^*)'Y] \end{aligned}$$

即$(a^*)'Y$也是$c'\beta$的无偏估计. 下面证明它是$c'\beta$的唯一的最小方差线性无偏估计.

设$b'Y$为$c'\beta$的任一无偏估计,从而只要令

$$b^* = \underset{\mu(X)}{\text{Proj}}\, b, \quad \tilde{b} = b - b^*$$

就同样地可以知道$(b^*)'Y$也是$c'\beta$的无偏估计. 由此有

$$\begin{aligned} 0 &= E[(a^*)'Y - (b^*)'Y] \\ &= (a^* - b^*)'X\beta \quad (\forall \beta \in R^p) \end{aligned}$$

所以

$$(a^* - b^*)'X = 0$$

注意到$a^* - b^* \in \mu(X)$,故$a^* = b^*$. 于是

$$\begin{aligned} \text{Var}(b'Y) &= b'\text{Cov}(Y,Y)b \\ &= b'(\sigma^2 I)b = \sigma^2 \|b\|^2 \\ &= \sigma^2(\|b^*\|^2 + \|\tilde{b}\|^2) \\ &= \sigma^2(\|a^*\|^2 + \|\tilde{b}\|^2) \\ &\geqslant \sigma^2 \|a^*\|^2 = \text{Var}((a^*)'Y) \end{aligned}$$

即$(a^*)'Y$是最小方差线性无偏估计. 如果我们进一步设$b'Y$也是最小方差的,即

$$\text{Var}(b'Y) = \text{Var}((a^*)'Y)$$

这时一定有

$$\|\tilde{b}\|^2 = 0, \quad \tilde{b} = 0$$

即

$$b = b^* = a^*, \quad b'Y = (a^*)'Y$$

即$(a^*)'Y$是$c'\beta$的唯一的最小方差线性无偏估计.

最后来证$(a^*)'Y = c'\hat{\beta}$,这里$\hat{\beta}$是β的任一个最小二乘估计.

由定理 3.1 的证明过程可以看出

$$\xi = X\hat{\beta} = \underset{\mu(X)}{\text{Proj}}\, Y$$

因 $a^* \in \mu(X)$，而 $Y-\xi \perp \mu(X)$，所以

$$(a^*)'(Y-X\hat{\beta})=0$$

$$(a^*)'Y=(a^*)'X\hat{\beta}$$

又由

$$c'\beta = E(a^*)'Y=(a^*)'X\beta, \quad \forall\, \beta \in R^p$$

知

$$(a^*)'X=c'$$

所以

$$(a^*)'Y=(a^*)'X\hat{\beta}=c'\hat{\beta}$$

证毕.

注 我们也可不用线性空间的投影理论，直接利用矩阵的基本运算给出定理 3.4 的证明.

设 $a'Y$ 是 $c'\beta$ 的任一无偏估计，则 $E(a'Y)=c'\beta$，即有 $a'X\beta=c'\beta$，这对一切 β 成立，故

$$a'X=c'$$

根据线性方程组的解的存在定理，有 β^* 满足 $X'X\beta^*=X'a$(因为增广矩阵的秩等于系数矩阵的秩)，令

$$a^*=X\beta^*$$

则 $X'a^*=X'a$，从而 $(a^*)'(a-a^*)=(\beta^*)'X'(a-a^*)=0$. 从 $a=a^*+(a-a^*)$知

$$\begin{aligned}\|a\|^2 &= \|a^*\|^2+\|a-a^*\|^2+2(a^*)'(a-a^*)\\ &= \|a^*\|^2+\|a-a^*\|^2\end{aligned}$$

我们指出$(a^*)'Y$ 也是 $c'\beta$ 的无偏估计，而且对于 β 的任一最小二乘估计 $\hat{\beta}$ 均有 $c'\hat{\beta}=(a^*)'Y$. 实际上，$E(a^*)'Y=(a^*)'X\beta=(\beta^*)'X'X\beta=(X'X\beta^*)'\beta=(X'a)'\beta=a'X\beta=c'\beta$，$(a^*)'Y=(X\beta^*)'Y=(\beta^*)'X'Y=(\beta^*)'X'X\hat{\beta}$(因为 $\hat{\beta}$ 满足正规方程(3.6))，故$(a^*)'Y=(a^*)'X\hat{\beta}=a'X\hat{\beta}=c'\hat{\beta}$.

另一方面，

$\text{Var}(a'Y)=a'\,\text{Cov}(Y,Y)\,a=\sigma^2\|a\|^2=\sigma^2(\|a^*\|^2+\|a-a^*\|^2)\geqslant \sigma^2\|a^*\|^2=\text{Var}((a^*)'Y)=\text{Var}(c'\hat{\beta})$.

此式中等号成立的充要条件是 $a=a^*$. 由此知 $c'\hat{\beta}$ 是 $c'\beta$ 的唯一的最小方差线性无偏估计.

定义 3.3 若 $\hat{\beta}$ 是 β 的最小二乘估计，则称 $c'\hat{\beta}$ 为 $c'\beta$ 的最小二乘估计.

注 如果 $c'\beta$ 可估，并且对线性模型(3.4)的假定 B 成立，则可以证明，$c'\hat{\beta}$ 是所有 $c'\beta$ 的无偏估计(不限于线性的)中方差最小的.

下面的定理把定理 3.2 的第三条结论推广到了 X 不满秩的情况.

定理 3.5 对线性模型(3.4)，设假定 A 成立，矩阵 X 的秩为 r，则

$$\hat{\sigma}^2=\frac{Q(\hat{\beta})}{n-r}$$

是 σ^2 的无偏估计.

证明：设 $\xi_1,\xi_2,\cdots,\xi_r$ 是 $\mu(X)$ 的一组标准正交基，添加 ξ_{r+1}，$\xi_{r+2},\cdots,\xi_n$ 使之成为 R^n 的标准正交基. 记

$$V_1=(\xi_1,\xi_2,\cdots,\xi_r)$$

$$V_2=(\xi_{r+1},\xi_{r+2},\cdots,\xi_n)$$

$$V=(V_1,V_2)$$

令 $Z=V'Y$，则

$$EZ=V'X\beta=\begin{pmatrix}V'_1\\V'_2\end{pmatrix}X\beta=\begin{pmatrix}V'_1X\beta\\0\end{pmatrix}$$

$$\mathrm{Cov}(Z,Z)=\sigma^2V'V=\sigma^2I$$

又 Y 可用 Z 表示，得

$$Y=VZ=(\xi_1,\xi_2,\cdots,\xi_n)\begin{pmatrix}z_1\\z_2\\\vdots\\z_n\end{pmatrix}$$

$$=\sum_{i=1}^n z_i\xi_i$$

从而

$$\xi \xlongequal{\mathrm{d}} \mathop{\mathrm{Proj}}_{\mu(X)} Y = \sum_{i=1}^{r} z_i \xi_i$$

$$Y - \xi = \sum_{i=r+1}^{n} z_i \xi_i$$

所以

$$Q(\hat{\beta}) = \|Y - \xi\|^2 = \sum_{i=r+1}^{n} z_i^2$$

$$\begin{aligned} EQ(\hat{\beta}) &= \sum_{i=r+1}^{n} E z_i^2 \\ &= \sum_{i=r+1}^{n} [\mathrm{Var}(z_i) + (E z_i)^2] \\ &= \sigma^2 (n - r) \end{aligned}$$

证毕.

如果假定 B 成立,可以进一步证明 $\hat{\sigma}^2$ 是 σ^2 的最小方差无偏估计.

(四) 带约束的线性模型的参数估计

在前面的讨论中我们对线性模型(3.4)中的参数 β 未作任何限制,因而对最小二乘估计的取值也没有限制. 在有些问题里,参数 β 满足一些给定的条件,要求估计量也满足同样的条件,现在就来研究这个问题.

例 3.3(平滑问题) 在天文测量中对空间中 3 个点构成的三角形 ABC 的 3 个内角 $\theta_1, \theta_2, \theta_3$ 进行测量,得到的观测值分别是 y_1, y_2, y_3. 由于测量是有误差的,所以就要找出 $\theta_1, \theta_2, \theta_3$ 的估计值. 当然 y_1, y_2, y_3 分别是 $\theta_1, \theta_2, \theta_3$ 的估计,但由于 $\theta_1 + \theta_2 + \theta_3 = \pi$,而 $y_1 + y_2 + y_3$ 不一定为 π,所以有必要找更好的估计 $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ 使 $\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 = \pi$. 设

$$y_1 = \theta_1 + e_1$$

$$y_2 = \theta_2 + e_2$$

$$y_3 = \theta_3 + e_3$$

e_1, e_2, e_3 为误差，我们假定测量无系统偏差且各次测量相互独立，各次测量的精度相等，于是

$$Ee_i = 0, \quad Ee_i e_j = \sigma^2 \delta_{ij} \quad (i, j = 1, 2, 3)$$

这是一个线性模型，问题是如何找到满足条件 $\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 = \pi$ 的估计量.

我们先讲一般理论，然后回到这个例子.

对线性模型(3.4)，设参数 β 满足约束

$$H\beta = r_0 \tag{3.12}$$

这里 H 是给定的 $s \times p$ 矩阵，r_0 为已知向量，且恒假定方程(3.12)有解. 如果要求 β 的任何估计 b 也满足

$$Hb = r_0$$

那么这样的估计问题就称为参数带(线性)约束的估计问题.

定义 3.4 记

$$\Theta_0 = \{\beta : H\beta = r_0\} \tag{3.13}$$

若估计量 $\hat{\beta}$ 满足

(1) $\hat{\beta} \in \Theta$;

(2) $\|Y - X\hat{\beta}\|^2 = \min\limits_{\beta \in \Theta_0} \|Y - X\beta\|^2$,

则称 $\hat{\beta}$ 是 β 的带约束条件(3.12)的最小二乘估计，简称约束最小二乘估计，也记 $\hat{\beta}$ 为 $\hat{\beta}_H$.

约束最小二乘估计永远存在. 有两种方法可以求约束最小二乘估计.

(1) 消去多余参数法

消去多余参数法的思想是从(3.12)中解出多余参数(即把它们用无约束的参数表示出来)代入(3.4)中用一般最小二乘法进行估计.

不失一般性，我们假设矩阵 H 的 s 行是线性无关的(这样应有 $s < p$，否则参数可直接由(3.12)解出)且头 s 列也是线性无关的. 记

$$H = (H_1, H_2)$$

其中 H_1 为 H 的前 s 列. 记

$$\beta=\begin{pmatrix}\beta^{(1)}\\ \beta^{(2)}\end{pmatrix},\quad \beta^{(1)}=\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_s\end{pmatrix}$$

设 $\beta\in\Theta_0$,则

$$H\beta=H_1\beta^{(1)}+H_2\beta^{(2)}=r_0$$

$$\beta^{(1)}=H_1^{-1}r_0-H_1^{-1}H_2\beta^{(2)}$$

这样模型(3.4)化为

$$Y=X\left[\begin{pmatrix}H_1^{-1}r_0\\ 0\end{pmatrix}+\begin{pmatrix}-H_1^{-1}H_2\\ I\end{pmatrix}\beta^{(2)}\right]+e$$

令

$$\widetilde{Y}=Y-X\begin{pmatrix}H_1^{-1}r_0\\ 0\end{pmatrix}$$

$$\widetilde{X}=X\begin{pmatrix}-H_1^{-1}H_2\\ I\end{pmatrix}$$

则

$$\widetilde{Y}=\widetilde{X}\beta^{(2)}+e$$

现在 $\beta^{(2)}$ 的取值已无约束,可由正规方程

$$\widetilde{X}'\widetilde{X}\hat{\beta}^{(2)}=\widetilde{X}'\widetilde{Y}$$

解出 $\beta^{(2)}$ 的最小二乘估计 $\hat{\beta}^{(2)}$,令

$$\hat{\beta}^{(1)}=H_1^{-1}r_0-H_1^{-1}H_2\hat{\beta}^{(2)}$$

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}^{(1)}\\ \hat{\beta}^{(2)}\end{pmatrix}$$

显然 $\hat{\beta}$ 满足约束条件,且由

$$\|Y-X\hat{\beta}\|^2=\|\widetilde{Y}-\widetilde{X}\hat{\beta}^{(2)}\|^2$$

易见 $\hat{\beta}$ 确是约束最小二乘估计.

消去多余参数法的缺点是要求逆矩阵,有时比较麻烦.

(2) 拉格朗日乘子法

约束最小二乘估计还可以用分析学中求条件极值的拉格朗日乘子法来求. 我们用代数方法给出下列定理.

定理 3.6 为了 $\hat{\beta}$ 是线性模型(3.4)的带约束条件(3.12)的最小二乘估计,必须且只须存在 λ 使 $\hat{\beta}$ 满足

$$\begin{cases} X'X\hat{\beta}-H'\lambda=X'Y \\ H\hat{\beta}=r_0 \end{cases} \tag{3.14}$$

证明:充分性. 设有 λ 使(3.14)成立,β 为任意的满足 $H\beta=r_0$ 的向量. 则

$$\begin{aligned}\|Y-X\beta\|^2&=\|(Y-X\hat{\beta})+X(\hat{\beta}-\beta)\|^2\\&=\|Y-X\hat{\beta}\|^2+\|X(\hat{\beta}-\beta)\|^2\\&\quad+2(Y-X\hat{\beta})'X(\hat{\beta}-\beta)\end{aligned}$$

注意到 $\hat{\beta}$ 和 β 都满足约束条件从而 $H(\hat{\beta}-\beta)=0$,又利用(3.14)可得

$$\begin{aligned}(Y-X\hat{\beta})'X(\hat{\beta}-\beta)&=(X'Y-X'X\hat{\beta})'(\hat{\beta}-\beta)\\&=-\lambda'H(\hat{\beta}-\beta)=0\end{aligned}$$

故

$$\begin{aligned}\|Y-X\beta\|^2&=\|Y-X\hat{\beta}\|^2+\|X(\hat{\beta}-\beta)\|^2\\&\geqslant\|Y-X\hat{\beta}\|^2\end{aligned}$$

即 $\hat{\beta}$ 是 β 的约束最小二乘估计.

必要性. 设 $\hat{\beta}$ 是满足约束条件(3.12)的 β 的最小二乘估计,当然(3.14)的第二个方程满足. 下面证明必有 λ 使得 $\hat{\beta}$ 满足(3.14)的第一个方程.

设 b 是满足方程 $Hb=0$ 的任一向量,令

$$\hat{\beta}=\beta+tb$$

t 为实数,则

$$H\beta=r_0$$

由约束最小二乘定义必有

$$\|Y-X\beta\|^2\geqslant\|Y-X\hat{\beta}\|^2$$

即有

$$\begin{aligned}&\|Y-X\beta\|^2-\|Y-X\hat{\beta}\|^2\\&=\|X(\hat{\beta}-\beta)\|^2+2(Y-X\hat{\beta})'X(\hat{\beta}-\beta)\\&=t^2\|Xb\|^2+2t(Y-X\hat{\beta})'Xb\geqslant 0\end{aligned}$$

由 t 的任意性及二次函数性质知

$$(Y-X\hat{\beta})'Xb=0$$
$$(X'Y-X'X\hat{\beta})'b=0$$

由于 b 是垂直于 $\mu(H')$（H 的各行构成的线性子空间）的任意向量，所以

$$X'X\hat{\beta}-X'Y\in\mu(H')$$

即存在 λ 使

$$X'X\hat{\beta}-X'Y=H'\lambda$$

证毕.

我们可以指出：

(1) 约束最小二乘估计永远存在，且存在性不依赖于假定 A 或假定 B.

(2) 对于带约束的最小二乘估计，也可以建立相应的高斯-马尔科夫定理.

例 3.3(续) 因只有 3 个参数，我们可用消去多余参数法来找出 $\theta_1,\theta_2,\theta_3$ 的优良估计. 令

$$z_1=y_1,\quad z_2=y_2,\quad z_3=y_3-\pi$$

则有

$$\begin{aligned}z_1&=\theta_1+e_1\\z_2&=\theta_2+e_2\\z_3&=-\theta_1-\theta_2+e_3\end{aligned}$$

即

$$\begin{pmatrix}z_1\\z_2\\z_3\end{pmatrix}=\begin{pmatrix}1&0\\0&1\\-1&-1\end{pmatrix}\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix}+\begin{pmatrix}e_1\\e_2\\e_3\end{pmatrix}$$

最小二乘估计为

$$\begin{pmatrix}\hat{\theta}_1\\\hat{\theta}_2\end{pmatrix}=\frac{1}{3}\begin{pmatrix}2z_1-z_2-z_3\\-z_1+2z_2-z_3\end{pmatrix}$$

$$
= \begin{bmatrix} y_1 - \frac{1}{3}(y_1 + y_2 + y_3 - \pi) \\ y_2 - \frac{1}{3}(y_1 + y_2 + y_3 - \pi) \end{bmatrix}
$$

又

$$
\hat{\theta}_3 = \pi - \hat{\theta}_1 - \hat{\theta}_2 = y_3 - \frac{1}{3}(y_1 + y_2 + y_3 - \pi)
$$

总之有

$$
\hat{\theta}_i = y_i - \frac{1}{3}(y_1 + y_2 + y_3 - \pi) \quad (i = 1, 2, 3)
$$

易见

$$
E\hat{\theta}_i = \theta_i
$$

$$
\mathrm{Var}(\hat{\theta}_i) = \frac{2}{3}\sigma^2 < \sigma^2 = \mathrm{Var}(y_i) \quad (i = 1, 2, 3)
$$

这表明用 $\hat{\theta}_i$ 估计 θ_i 比用单个 y_i 估计 θ_i 有较小的方差. 这是平滑的好处.

这个问题如果用拉格朗日乘子法来解，则方程可以写成

$$
\begin{cases} \theta_1 - \lambda = y_1 \\ \theta_2 - \lambda = y_2 \\ \theta_3 - \lambda = y_3 \\ \theta_1 + \theta_2 + \theta_3 = \pi \end{cases}
$$

同样也可解出 $\hat{\theta}_i = y_i - \frac{1}{3}(y_1 + y_2 + y_3 - \pi)$.

(五) 关于最小二乘估计的讨论

作为本节的结尾，我们对最小二乘估计的优缺点要说几句话. 迄今为止，在对线性模型

$$
Y = X\beta + e
$$

的未知参数 β 进行估计时，广泛使用最小二乘估计. 当 X 满秩时，所得的最小二乘估计是无偏的而且是所有线性无偏估计中方差最小的，计算也不复杂，只需解线性方程组. 这些都是最小二乘估计

的优点. 但是,并不是在任何情况下,最小二乘估计都好.

(1) 当行列式 $|X'X|$ 接近零时,最小二乘估计 $\hat{\beta}$ 很差,取值很不稳定,这是由于

$$\mathrm{Cov}(\hat{\beta},\hat{\beta})=\sigma^2(X'X)^{-1}$$

碰到 $|X'X|$ 接近零的情况时(当 X 的各列近似地线性相关时就会出现这种情况),就得考虑其他的估计方法. 一个重要的估计是

$$\tilde{\beta}=(X'X+\lambda I)^{-1}X'Y$$

其中 λ 是正参数(当 $\lambda=0$ 时 $\tilde{\beta}$ 即最小二乘估计). 这种估计通称岭估计,是由 Hoerl 和 Kennard 于 1970 年提出的,关于其中的参数 λ 的选取现代有许多研究成果,这里就不细述了.

(2) 最小二乘估计是基于最小化残差平方和而得到的估计法,它可能受到个别数据的较大影响. 换句话说,稳健性不够好. 30 多年来,稳健统计的思想和方法得到了很大发展. 比最小二乘估计稳健的估计有最小一乘估计,它的定义是:若 $\beta^*=(\beta_1^*,\beta_2^*,\cdots,\beta_p^*)'$ 使得

$$\sum_{i=1}^{n}|y_i-(x_{i1}\beta_1+x_{i2}\beta_2+\cdots+x_{ip}\beta_p)|$$

达到最小,则称 β^* 是 β 的最小一乘估计. 求最小一乘估计要用到线性规划的方法,其统计性质的研究也比较复杂,已有不少研究成果.

§4 线性模型的假设检验

(一) 参数线性相关性的检验

给定线性模型

$$Y=X\beta+e \tag{4.1}$$

其中 X 是已知的 $n\times p$ 矩阵($n>p$),β 是 p 维未知参数向量,Y 是 n 维观测向量,e 是随机项. 本节恒假设假定 B 成立,即

$$e\sim N(0,\sigma^2 I) \tag{4.2}$$

以下以 r 表示 X 的秩，ξ 表示 EY.

考虑检验问题

$$H_0: H\beta=0 \leftrightarrow H_a: H\beta\neq 0 \tag{4.3}$$

其中 H 为 $s\times p$ 矩阵.

令

$$W=\mu(X)=\{\eta=X\beta: \beta\in R^p\}\subset R^n$$
$$W_0=\{\eta=X\beta: H\beta=0,\beta\in R^p\}\subset W$$

则假设 H_0 等价于 $\xi\in W_0$. 设 W_0 的维数为 q 且 $q<r$.

我们用广义似然比来构造检验(4.3)的检验统计量. 记

$$\Theta=\{(\xi,\sigma^2): \xi\in W,\sigma^2>0\}$$
$$\Theta_0=\{(\xi,\sigma^2): \xi\in W_0,\sigma^2>0\}$$

易见 Y 的分布密度为

$$p(Y;\xi,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\|Y-\xi\|^2\right\}$$

这也是参数(ξ,σ^2)的似然函数. 不难看出，要使这个函数达到最大值，必须且只须选 ξ 使 $\|Y-\xi\|^2$ 最小，再选 σ^2 使表达式达到最大值. 记

$$\hat{\xi}=\operatorname*{Proj}_{W} Y,\quad \hat{\xi}_0=\operatorname*{Proj}_{W_0} Y$$

易知 $p(Y,\xi,\sigma^2)$在 $\xi=\hat{\xi},\sigma^2=\frac{1}{n}\|Y-\hat{\xi}\|^2$ 时达到最大值 $\sup\limits_{(\xi,\sigma^2)\in\Theta} p(Y,\xi,\sigma^2)$，在 $\xi=\hat{\xi}_0,\sigma^2=\frac{1}{n}\|Y-\hat{\xi}_0\|^2$ 时达到最大值 $\sup\limits_{(\xi,\sigma^2)\in\Theta_0} p(Y,\xi,\sigma^2)$.

于是广义似然比

$$\lambda=\frac{\sup\limits_{\Theta} p(Y,\xi,\sigma^2)}{\sup\limits_{\Theta_0} p(Y,\xi,\sigma^2)}$$

$$=\left\{\frac{\|Y-\hat{\xi}_0\|^2}{\|Y-\hat{\xi}\|^2}\right\}^{\frac{n}{2}}$$

注意到 $Y-\hat{\xi}\perp W$，而 $\hat{\xi}-\hat{\xi}_0\in W$，故 $Y-\hat{\xi}\perp\hat{\xi}-\xi_0$. 从而有

$$\|Y-\hat{\xi}_0\|^2=\|Y-\hat{\xi}\|^2+\|\hat{\xi}-\hat{\xi}_0\|^2$$

所以

$$\lambda=\left(1+\frac{\|\hat{\xi}-\hat{\xi}_0\|^2}{\|Y-\hat{\xi}\|^2}\right)^{\frac{n}{2}}$$

$$\lambda>\lambda_0 \Leftrightarrow \frac{\|\hat{\xi}-\hat{\xi}_0\|^2}{\|Y-\hat{\xi}\|^2}>\lambda_1$$

令

$$F=\frac{\|\hat{\xi}-\hat{\xi}_0\|^2/(r-q)}{\|Y-\hat{\xi}\|^2/(n-r)}$$

对否定域$\{\lambda>\lambda_0\}$必有 C 使$\{F>C\}=\{\lambda>\lambda_0\}$. 因此 F 可用作检验统计量. 下面的定理给出了零假设下 F 的分布.

定理 4.1 对线性模型(4.1),设(4.2)成立,则在假设 H_0 成立时,统计量 F 的分布为 $F(r-q,n-r)$.

证明:在 R^n 中取标准正交基 $\xi_1,\xi_2,\cdots,\xi_n$(均为列向量)使得 $\xi_1,\xi_2,\cdots,\xi_q$ 为 W_0 的基,$\xi_1,\xi_2,\cdots,\xi_r$ 为 W 的基. 记

$$V=(\xi_1,\xi_2,\cdots,\xi_n)$$

$$V_1=(\xi_1,\xi_2,\cdots,\xi_q)$$

$$V_2=(\xi_{q+1},\xi_{q+2},\cdots,\xi_n)$$

则

$$V=(V_1,V_2)$$

令

$$Z=\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}=V'Y$$

则

$$Y=VZ=\sum_{i=1}^{n} z_i\xi_i$$

$$\hat{\xi}_0 = \sum_{i=1}^{q} z_i \xi_i, \quad \hat{\xi} = \sum_{i=1}^{r} z_i \xi_i$$

$$\hat{\xi} - \hat{\xi}_0 = \sum_{i=q+1}^{r} z_i \xi_i$$

故

$$F = \frac{\left\| \sum_{i=q+1}^{r} z_i \xi_i \right\|^2 / (r-q)}{\left\| \sum_{i=r+1}^{n} z_i \xi_i \right\|^2 / (n-r)} = \frac{\sum_{i=q+1}^{r} z_i^2 / (r-q)}{\sum_{i=r+1}^{r} z_i^2 / (n-r)}$$

又由 $Z=V'Y$ 得

$$Z \sim N(V'\xi, \sigma^2 V'V),$$

其中 $V'V=I$,而

$$V'\xi = \begin{pmatrix} V'_1 \\ V'_2 \end{pmatrix} \xi = \begin{pmatrix} V'_1 \xi \\ \Theta \end{pmatrix}$$

(因为在 H_0 下应有 $\xi \in W_0$,而 V_2 的各列垂直于 W_0).

于是有

$$Z \sim N\left(\begin{pmatrix} V'_1 \xi \\ 0 \end{pmatrix}, \sigma^2 I \right)$$

这说明 $z_1, z_2, \cdots, z_n$ 独立,且 $z_{q+1}, z_{q+2}, \cdots, z_n$ 同为 $N(0, \sigma^2)$ 分布. 所以在 H_0 下有

$$\frac{1}{\sigma^2} \sum_{i=q+1}^{r} z_i^2 \sim \chi^2(r-q)$$

$$\frac{1}{\sigma^2} \sum_{i=r+1}^{n} z_i^2 \sim \chi^2(n-r)$$

且二者独立. 故在 H_0 下 F 服从 $F(r-q, n-r)$ 分布. 证毕.

注 我们可以不使用线性空间的投影理论，直接表述统计量 F. 设 $\hat{\beta}$ 为 β 的一个最小二乘估计，$\hat{\beta}_0$ 为 β 在约束条件 $H\beta=0$ 下的一个最小二乘估计，只要用 $X\hat{\beta}$ 和 $X\hat{\beta}_0$ 分别代替上文的 $\hat{\xi}$ 和 $\hat{\xi}_0$，不难验证上述推导仍然成立. 这时 $\|Y-\hat{\xi}\|^2=\|Y-X\hat{\beta}\|^2$ 是残差平方和 Q; $\|Y-\hat{\xi}_0\|^2=\|Y-X\hat{\beta}_0\|^2$ 是约束条件 $H\beta=0$ 下的残差平方和，常记为 Q_H. 于是 $\|\hat{\xi}-\hat{\xi}_0\|^2=Q_H-Q$, $F=\dfrac{(Q_H-Q)/(r-q)}{Q/(n-r)}$，这是常用的公式.

用 $\hat{\beta}$ 表示 β 的最小二乘估计，则 $\hat{\xi}=X\hat{\beta}$(见定理 3.1 的证明)，令

$$\hat{e}=Y-\hat{\xi}=Y-X\hat{\beta}$$

称 $\hat{e}$ 为残差，记

$$Q=\|\hat{e}\|^2=\sum_{i=1}^{n}\{y_i-(x_{i1}\hat{\beta}_1+x_{i2}\hat{\beta}_2+\cdots+x_{ip}\hat{\beta}_p)\}^2$$

它是残差平方和，$X\hat{\beta}$ 和 Q 之间有如下关系：

系 4.1 对模型(4.1)，设条件(4.2)成立，$\hat{\beta}$ 为 β 的最小二乘估计，则 $X\hat{\beta}$ 与残差 $\hat{e}$ 独立，从而 $X\hat{\beta}$ 与残差平方和 Q 独立，且

$$Q/\sigma^2\sim\chi^2(n-r)$$

证明：不论假设 H_0 是否成立，在定理 4.1 证明中总有 $Z\sim N(V'\xi,\sigma^2 I)$，从而独立性得证. 又 $\xi\in W$，故 ξ 与 $\xi_{r+1},\xi_{r+2},\cdots,\xi_n$ 正交，因此

$$V'\xi=\begin{pmatrix}\xi'_1\\ \vdots\\ \xi'_r\\ \xi_{r+1}\\ \vdots\\ \xi_n\end{pmatrix}\xi=\begin{pmatrix}\xi'_1\xi\\ \vdots\\ \xi'_r\xi\\ 0\\ \vdots\\ 0\end{pmatrix},$$

故 $z_{r+1},z_{r+2},\cdots,z_n$ 独立同 $N(0,\sigma^2)$ 分布. 于是可知

$$Q/\sigma^2=\frac{1}{\sigma^2}\sum_{i=r+1}^{n}z_i^2\sim\chi^2(n-r)$$

证毕.

(二) 参数线性组合的检验与置信区间

上面证明了 $X\hat{\beta}$ 与 Q 独立，$\hat{\beta}$ 与 Q 是否相互独立呢？从下面的讨论中将可以看到，当 X 满秩时，$\hat{\beta}$ 与 Q 独立. 对 β 的可估线性组合有如下的一般结论：

定理 4.2 对模型(4.1)，设条件(4.2)成立，$c'\beta$ 是 β 的可估线性组合，$\hat{\beta}$ 是 β 的最小二乘估计，Q 为残差平方和，则 $c'\hat{\beta}$ 与 Q 相互独立.

证明：因 $c'\beta$ 可估，由定理 3.4 的证明过程可知有 $a^* \in \mu(X)$ 使

$$c'\hat{\beta}=a^* Y$$

在 R^n 中取标准正交基 $\xi_1,\xi_2,\cdots,\xi_n$ 使得 $\xi_1,\xi_2,\cdots,\xi_r$ 构成 $\mu(X)$ 的基，记

$$V_1=(\xi_1,\xi_2,\cdots,\xi_r),\quad V_2=(\xi_{r+1},\cdots,\xi_n)$$
$$V=(V_1,V_2)$$

令

$$Z=V'Y$$

则

$$Z\sim N(V'X\beta,\sigma^2 I)$$

故 $z_1,z_2,\cdots,z_n$ 相互独立，又

$$\begin{aligned} c'\hat{\beta} &=(a^*)'Y=(a^*)'VZ=(a^*)'(V_1,V_2)Z \\ &=((a^*)'V_1,0)Z \\ &=\varphi(z_1,z_2,\cdots,z_r) \end{aligned} \tag{4.4}$$

而残差 $\hat{e}=\sum_{i=r+1}^{n} z_i\xi_i$，$Q=\|\hat{e}\|^2=\sum_{i=r+1}^{n} z_i^2$，所以 $c'\hat{\beta}$ 与 Q 相互独立.

证毕.

系 4.2 在定理 4.2 条件下，如果 X 满秩，则 $\hat{\beta}$ 与 Q 独立，且

$$\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1}),\quad Q/\sigma^2\sim\chi^2(n-p)$$

证明：前一节我们已经说过，X 满秩则任意的 $c'\beta$ 都是可估的. 从(4.4)知 $\hat{\beta}_i$ 只与 $z_1,\cdots,z_r$ 有关，而 Q 只依赖于 $z_{r+1},\cdots,z_n$，

故 $\hat{\beta}$ 与 Q 独立. 因 $\hat{\beta}=(X'X)^{-1}X'Y$ 是 Y 的线性组合,而 Y 服从 n 维正态分布,又 $E\hat{\beta}=\beta$, $\mathrm{Cov}(\hat{\beta})=\sigma^2(X'X)^{-1}$,故 $\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1})$. 从定理 4.1 的系知 $Q/\sigma^2\sim\chi^2(n-p)$.

证毕.

定义 4.1 设 $a'Y$ 是 $c'\beta$ 的无偏估计且 $a\in\mu(X)$,则称 a 为 c 的伴随元.

从定理 3.4 的证明过程可知,若 $c'\beta$ 可估,则 c 的伴随元存在唯一,当 X 满秩时,c 的伴随元为

$$a=X(X'X)^{-1}c$$

(这时用 $c'\hat{\beta}=c'(X'X)^{-1}X'Y=a'Y$ 来估计 $c'\beta$).

定理 4.3 对线性模型(4.1),设条件(4.2)成立,并设 $c'\beta$ 可估($c\neq 0$),则

$$\frac{c'(\hat{\beta}-\beta)}{\hat{\sigma}\|a\|}\sim t(n-r)$$

其中 a 为 c 的伴随元,r 为 X 的秩,$\hat{\sigma}=\sqrt{\dfrac{Q}{n-r}}$.

证明:由定理 3.4 的证明过程可知 $c'\hat{\beta}=a'Y$,于是

$$\begin{aligned}\mathrm{Var}(c'\hat{\beta})&=\mathrm{Var}(a'Y)=a'\mathrm{Cov}(Y,Y)a\\&=\|a\|^2\sigma^2\end{aligned}$$

故

$$c'\hat{\beta}\sim N(c'\beta,\sigma^2\|a\|^2)$$

又由定理 4.2 知 $c'\hat{\beta}$ 与 Q 独立,又

$$Q/\sigma^2\sim\chi^2(n-r)$$

故

$$\frac{c'\hat{\beta}-c'\beta}{\|a\|}\Big/\sqrt{\frac{Q}{n-r}}\sim t(n-r)$$

即

$$\frac{c'(\hat{\beta}-\beta)}{\hat{\sigma}\|a\|}\sim t(n-r)$$

证毕.

系 4.3 在定理 4.3 的条件下,若 X 满秩,则

$$\frac{c'(\hat{\beta}-\beta)}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}} \sim t(n-p)$$

证明:这时 $a=X(X'X)^{-1}c$,于是

$$\|a\|^2=c'(X'X)^{-1}X'X(X'X)^{-1}c=c'(X'X)^{-1}c$$

证毕.

有了以上的理论,我们就可以进行假设检验和构造置信区间.

(1) 若 $c'\beta$ 可估,要检验

$$H_0: c'\beta=r_0$$

则可用统计量

$$t=\frac{c'\hat{\beta}-r_0}{\hat{\sigma}\|a\|} \quad (a \text{ 为 } c \text{ 的伴随元})$$

在 H_0 下 $t\sim t(n-r)$(r 为 X 的秩),取 λ 使得

$$P(|t|>\lambda \mid H_0)=\alpha$$

则在 α 水平下检验的否定域为 $\{|t|>\lambda\}$.

$H_0: c'\beta\leqslant r_0$,$H_0: c'\beta\geqslant r_0$ 等假设也可类似地使用 t 统计量进行检验.

(2) 若 $c'\beta$ 可估,我们可以求出 $c'\beta$ 的置信区间. 因

$$t=\frac{c'\hat{\beta}-c'\beta}{\hat{\sigma}\|a\|} \sim t(n-r)$$

取 λ 使 $P(|t|>\lambda)=\alpha$,则 $c'\beta$ 的 $1-\alpha$ 置信水平的置信区间为

$$[c'\hat{\beta}-\lambda\|a\|\hat{\sigma},\ c'\hat{\beta}+\lambda\|a\|\hat{\sigma}]$$

例 4.1 我们来检验一下例 2.2 中模型的相关性. 这里

$$X=(22,68,108,\cdots,1\,100)'$$

$$Y=(0.75,2.4,3.2,\cdots,38.0)'$$

$$X'X=2\,685\,141, \quad Y'Y=3\,376.312\,5$$

最小二乘估计 $\hat{\beta}=\hat{H}=0.035\,44$,来检验

$$H_0: H=0$$

即取 $c=1$,检验 $H_0: c'\beta=0(\beta=H)$. 易见 $c=1$ 的伴随元为 $a=X(X'X)^{-1}$, $\|a\|=\sqrt{X'X}$,取统计量

$$t=\frac{\hat{\beta}-0}{\hat{\sigma}\|a\|}$$

在 H_0 下 $t\sim t(11-1)$. 计算得

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1}\|Y-X\hat{\beta}\|^2 \\ &= \frac{1}{10}(Y'Y-X'X\hat{H}^2) \\ &=0.379\,26 \\ \hat{\sigma} &=0.615\,8 \\ t &=94.31\end{aligned}$$

查得 $t(10)$分布的 1－0.005 分位点才为 3.169,可见应拒绝 H_0. 星团距离与速度之间确有线性相关关系.

用上面的 $\hat{\sigma}$, $\|a\|$还可算出 H 的 99%置信水平的置信区间为 $|H-\hat{H}|\leqslant 0.001\,2$,即 $H\in[0.034\,25, 0.036\,63]$.

§5 回 归 分 析

我们本节利用§3、§4 中建立的线性模型的一般理论来研究§1 中提出的回归分析问题. 我们的出发点是因变量 y 与自变量 $x_1, x_2, \cdots, x_p$ 的下列数据

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{matrix} \qquad (5.1)$$

需要解决以下问题:

(1) y 与 $x_1, x_2, \cdots, x_p$ 是否有下列线性关系:

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+e \qquad (5.2)$$

其中 $Ee=0$,即是否 $\mathrm{Var}(e)$ 足够小使得下列近似式

$$y \doteq \beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

有意义.

(2) 若 y 与 $x_1,x_2,\cdots,x_p$ 有(5.2)的关系,如何找出 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$.

(3) 如何利用回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_p x_p \tag{5.3}$$

进行预测和控制.

我们先回答问题(2)、(3),再来研究(1).

(一) 回归参数估计

我们用 §3 中介绍的最小二乘法来估计回归参数. 记

$$X=(E,C)$$

其中

$$E=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{n\times 1},\quad C=(x_{ij})_{n\times p}$$

并记

$$\beta=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix},\quad e=\begin{pmatrix}e_1\\e_2\\\vdots\\e_n\end{pmatrix},\quad Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}$$

则由(5.1)和(5.2)有

$$Y=X\beta+e \tag{5.4}$$

在回归分析中,我们恒假定数据(5.1)使 X 的各列线性无关,即 X 的秩为 $p+1$. 由 §3 知 β 的最小二乘估计为

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{5.5}$$

我们要找一个更便于计算的公式. 记

$$l_{ij}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_{\cdot i})(x_{kj}-\bar{x}_{\cdot j})$$

$$i,j = 1,2,\cdots,p$$

$$l_{iy} = \sum_{k=1}^{n}(x_{ki} - \overline{x}_{\cdot i})(y_k - \overline{y}), \quad i = 1,2,\cdots,p$$

$$l_{yy} = \sum_{k=1}^{n}(y_k - \overline{y})^2$$

$$L = (l_{ij})_{p\times p}, \quad L_{xy} = (l_{iy})_{p\times 1}$$

$$\overline{x} = \begin{pmatrix} \overline{x}_{\cdot 1} \\ \overline{x}_{\cdot 2} \\ \vdots \\ \overline{x}_{\cdot p} \end{pmatrix}$$

其中

$$\overline{x}_{\cdot i} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}, \quad i = 1,2,\cdots,p$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

这里 L 叫做相关矩阵. 易见

$$l_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj} - n\overline{x}_{\cdot i}\overline{x}_{\cdot j}$$

$$L = C'C - n\overline{x}\,\overline{x}'$$

$$X'X = \begin{pmatrix} n & E'C \\ C'E & C'C \end{pmatrix} = \begin{pmatrix} n & n\overline{x}' \\ n\overline{x} & L + n\overline{x}\,\overline{x}' \end{pmatrix}$$

$$l_{iy} = \sum_{k=1}^{n} x_{ki}y_i - n\overline{x}_{\cdot i}\overline{y}$$

$$X'Y = \begin{pmatrix} E'Y \\ C'Y \end{pmatrix} = \begin{pmatrix} n\overline{y} \\ L_{xy} + n\overline{y}\cdot\overline{x} \end{pmatrix}$$

用矩阵四块求逆公式可得

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{x}'L^{-1}\overline{x} & -\overline{x}'L^{-1} \\ -L^{-1}\overline{x} & L^{-1} \end{pmatrix}$$

从而

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \overline{y} - \overline{x}'L^{-1}L_{xy} \\ L^{-1}L_{xy} \end{pmatrix}$$

所以 $\beta_0, \beta_1, \cdots, \beta_p$ 的最小二乘估计为

$$\hat{\beta}_0 = \overline{y} - \sum_{i=1}^{p} \hat{\beta}_i \overline{x}_{\cdot i}$$

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = L^{-1} \begin{pmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{py} \end{pmatrix} = L^{-1}L_{xy} \tag{5.6}$$

残差平方和

$$\begin{aligned} Q &= \|Y - X\hat{\beta}\|^2 = Y'Y - Y'X\hat{\beta} \\ &= l_{yy} + n\overline{y}^2 - \begin{pmatrix} n\overline{y} \\ L_{xy} + n\overline{y} \cdot \overline{x} \end{pmatrix}' \hat{\beta} \\ &= l_{yy} - L'_{xy}L^{-1}L_{xy} \\ &= l_{yy} - \sum_{j=1}^{p} l_{jy}\hat{\beta}_j \end{aligned} \tag{5.7}$$

误差标准差估计为

$$\hat{\sigma} = \sqrt{\frac{Q}{n-p-1}} \tag{5.8}$$

由定理 3.2，如果设(5.4)中的随机项 e 满足

$$Ee = 0, \quad Eee' = \sigma^2 I$$

则最小二乘估计是无偏估计，且

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}, \hat{\beta}) &= \sigma^2 (X'X)^{-1} \\ &= \sigma^2 \begin{pmatrix} \frac{1}{n} + \overline{x}'L^{-1}\overline{x} & -\overline{x}'L^{-1} \\ -L^{-1}\overline{x} & L^{-1} \end{pmatrix} \end{aligned} \tag{5.9}$$

$$E\hat{\sigma}^2 = \sigma^2$$

（二）预测

有了回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

则对一组自变量值 $x_1=x_1^0, x_2=x_2^0, \cdots, x_p=x_p^0$，可得 y 的预测值

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0$$

易见

$$E\hat{y}=\beta_0+\beta_1 x_1^0+\cdots+\beta_p x_p^0=Ey$$

即 $\hat{y}$ 是期望值 $Ey=\beta_0+\beta_1 x_1^0+\cdots+\beta_p x_p^0$ 的无偏估计.

下面来找 Ey 的置信区间. 设(5.2)中随机项 $e\sim N(0,\sigma^2)$，则令

$$c=(1,x_1^0,\cdots,x_p^0)'$$

有

$$Ey=c'\beta$$

是一个线性可估量(仍假设(5.4)中的 X 满秩)，由定理 4.3 的系可知

$$\frac{c'\hat{\beta}-c'\beta}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}}\sim t(n-p-1)$$

记

$$x^0=(x_1^0,x_2^0,\cdots,x_p^0)'$$

则

$$c'(X'X)^{-1}c=\frac{1}{n}+(x^0-\overline{x})'L^{-1}(x^0-\overline{x})$$

因此

$$t\overset{\mathrm{d}}{=}\frac{(\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0)-c'\beta}{\hat{\sigma}\sqrt{\dfrac{1}{n}+(x^0-\overline{x})'L^{-1}(x^0-\overline{x})}}\sim t(n-p-1) \quad (5.10)$$

查 t 分布表知有 λ 满足 $P(|t|>\lambda)=\alpha$.

记

$$u=\lambda\hat{\sigma}\sqrt{\frac{1}{n}+(x^0-\overline{x})'L^{-1}(x^0-\overline{x})}$$

则 $Ey=c'\beta=\beta_0+\beta_1 x_1^0+\cdots+\beta_p x_p^0$ 的 $1-\alpha$ 置信区间为

$$[\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0-u,\ \hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0+u]$$

下面我们来求真值 y 的置信区间. y 的预测值为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0=c'\hat{\beta}$$

预测误差

$$\begin{aligned}\delta&=y-\hat{y}=y-c'\hat{\beta}\\&=(y-c'\beta)-c'(\hat{\beta}-\beta)\end{aligned}$$

注意到 y 是模型(5.2)在 $x_1=x_1^0,\cdots,x_p=x_p^0$ 处的一个新的观测值,所以我们仍假定对应的随机项

$$e_0=y-\beta_0-\beta_1 x_1^0-\cdots-\beta_p x_p^0=y-c'\beta$$

与前 n 个观测的随机项 $e_1,e_2,\cdots,e_n$ 独立且有相同的 $N(0,\sigma^2)$ 分布,这样就有

$$\begin{aligned}\operatorname{Var}(\delta)&=\sigma^2+\sigma^2 c'(X'X)^{-1}c\\&=\sigma^2\left[1+\frac{1}{n}+(x^0-\bar{x})'L^{-1}(x^0-\bar{x})\right]\end{aligned}$$

所以

$$\delta\sim N\left(0,\sigma^2\left[1+\frac{1}{n}+(x^0-\bar{x})'L^{-1}(x^0-\bar{x})\right]\right)$$

于是

$$t=\frac{y-(\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0)}{\hat{\sigma}\sqrt{1+\frac{1}{n}+(x^0-\bar{x})'L^{-1}(x^0-\bar{x})}}\sim t(n-p-1) \quad (5.11)$$

令

$$v=\lambda\hat{\sigma}\sqrt{1+\frac{1}{n}+(x^0-\bar{x})'L^{-1}(x^0-\bar{x})}$$

则真值 $y=\beta_0+\beta_1 x_1^0+\cdots+\beta_p x_p^0+e_0$ 的 $1-\alpha$ 置信区间为

$$[\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0-v,\hat{\beta}_0+\hat{\beta}_1 x_1^0+\cdots+\hat{\beta}_p x_p^0+v]$$

它比 Ey 的置信区间要大,因为 y 比 Ey 多了一个随机项 e_0.

应该注意的是(5.11)成立的条件是$(x_1^0,\cdots,x_p^0,y)$符合原来的回归模型,这要求观测$(x_1^0,\cdots,x_p^0,y)$时变量之间的关系没有大的改变. 因此(5.11)的 t 统计量可以用来检验假设

H_0:新得到的数据$(x_1^0,\cdots,x_p^0,y)$与建立回归方程所用数据来自同一线性回归模型

当 t 的绝对值大于某个临界值时拒绝 H_0，即观测$(x_1^0,\cdots,x_p^0;y)$时变量间的关系已有变化.

例 5.1 表 5.1 是英国 11 年的有执照汽车数与车祸次数的数据. 从数据中可以看出，汽车数增加时车祸数也相应地增加了. 我们对车祸数(y)和汽车数(x)来拟合线性回归式. 计算得

$$n=11,\quad \sum x_i=5\ 711,\quad \sum y_i=2\ 396$$

$$\sum x_i^2=3\ 134\ 543,\quad \sum x_iy_i=1\ 296\ 836,\quad \sum y_i^2=539\ 512$$

求得回归方程

$$\hat{y}=55.86+0.312\ 0x$$

相应的 F 统计量

$$F=132.0$$

自由度(1,9)的 F 分布的 0.99 分位点为 10.56，可见回归方程是显著的.

表 5.1 有关英国车祸次数与有执照汽车数的数据

年份	车祸次数 y(千次)	有执照汽车数 x(万辆)
1947	166	352
1948	153	373
1949	177	411
1950	201	441
1951	216	462
1952	208	490
1953	227	529
1954	238	577
1955	268	641
1956	268	692
1957	274	743

下面我们用得到的回归式进行预测. 应当强调指出的是，在

实际工作中，不能仅仅根据统计关系做简单的外推，因为得出已知统计关系的实际背景可能有变化. 在这个问题中，必须对与发生车祸有关的因素进行详细的研究，比如：公路和交通工具的发展状况，新增加的汽车的种类和性能以及人们如何使用这些汽车等等，这些与车祸有关的复杂因素是否没有大的变化.

相对于有 1 000 万辆汽车，即 $x^0=1\,000$，可得预测值

$$\hat{y}=367.85$$

即可能发生大约 367 850 次车祸. 易见这时(5.11)的 t 统计量可以写成

$$t=\frac{y-\hat{y}}{\hat{\sigma}\sqrt{1+\frac{1}{n}+(x^0-\bar{x})^2/l_{11}}}$$

由此可得车祸数的 95％预测区间为

$$367.85\pm(2.262)(11.18)\sqrt{1+\frac{1}{11}+\frac{(1\,000-519.18)^2}{169\,495.64}}$$

即 328 227 次到 407 473 次之间.

假设在下一年内，交通管理部门对违章驾驶进行了更严厉的处罚，汽车数是 800 万辆，发生了 270 千次车祸. 我们想知道汽车数与车祸数之间的关系是否发生了显著的变化. 设

$$x^0=800, \quad y=270$$

可以算出

$$\hat{y}=305.45$$

及 t 统计量值

$$t=\frac{y-\hat{y}}{\hat{\sigma}\sqrt{1+\frac{1}{n}+(x^0-\bar{x})^2/l_{11}}}=-2.54$$

考虑假设

H_0：车祸数与汽车数关系不变

H_1：车祸数减少了

可以用单侧否定域$\{t<\lambda_0\}$，9 个自由度的 t 分布的 0.05 分位数为

-1.833，0.01 分位数为 -2.821，因此水平 0.05 下认为车祸数确实减少了，水平 0.01 下不能否定原假设.

（三）控制

控制问题是要求因变量 y 以 $1-\alpha$ 的概率取值于区间 $[A,B]$，问自变量 $x_1,x_2,\cdots,x_p$ 应控制在什么范围内. 这是预测问题的反问题.

查 t 分布表得 $n-p-1$ 自由度时 $1-\frac{\alpha}{2}$ 分位数（即双侧 α 水平分位点）λ，则下列不等式以概率 $1-\alpha$ 成立：

$$(y-\hat{\beta}_0-\hat{\beta}_1x_1-\cdots-\hat{\beta}_px_p)^2$$
$$\leqslant\lambda^2\hat{\sigma}^2\left(1+\frac{1}{n}+(x-\bar{x})'L^{-1}(x-\bar{x})\right)$$

其中

$$x=(x_1,x_2,\cdots,x_p)',\quad \bar{x}=(\bar{x}_{\cdot 1},\bar{x}_{\cdot 2},\cdots,\bar{x}_{\cdot p})'$$

根据要求 $A\leqslant y\leqslant B$，从上述不等式就可以找出 $x_1,x_2,\cdots,x_p$ 的控制范围.

例 5.2 低钴定膨胀合金由铁、镍、钴、铜组成. 在控制杂质含量及一定的工艺条件下，其膨胀特性被合金成分所确定. 我国某课题组（1975 年）的研究任务就是：确定合适的合金成分，使得钴的用量尽量少，但使得合金的膨胀系数与瓷封材料的膨胀系数相当（在 5.5～8.0 之间，单位：10^{-6}/℃）.

这就是一个控制问题，首先要建立回归关系式. 设铜的百分含量为 x_1，镍的百分含量减去 30 后记为 x_2，钴的百分含量为 x_3，记 α_{300}，α_{500} 为 300 ℃及 500 ℃时合金的膨胀系数，它们都是 x_1,x_2,x_3 的函数，考虑到各种误差因素，设

$$\alpha_{300}=f_1(x_1,x_2,x_3)+\varepsilon_1$$
$$\alpha_{500}=f_2(x_1,x_2,x_3)+\varepsilon_2$$

只能根据试验数据去确定 f_1 和 f_2，也就是建立经验公式. 因为线性函数明显不符合要求，所以设想用三元二次多项式模型

$$f(x_1,x_2,x_3)=b_0+b_1x_1+b_2x_2+b_3x_3+ b_4x_1^2+b_5x_1x_2+b_6x_1x_3+ b_7x_2^2+b_8x_2x_3+b_9x_3^2$$

为了确定系数 $b_0,b_1,\cdots,b_9$，该课题组炼了 2 kg 钢锭 60 炉，其名义成分如下：

铜/%	0.5	1.5	3	5	7
镍/%	32	34	35.5	37	
钴/%	4	6	8		

全部组合为 5×4×3＝60 种．各炉的实际成分及膨胀系数见表5.2.

表 5.2　合金试验数据(部分)

编号	铜(%)	镍(%)	钴(%)	α_{300} (10^{-6})	α_{500} (10^{-6})
1	0.55	32.09	3.72	5.76	10.44
2	0.50	31.66	5.38	4.12	9.23
3	0.51	31.64	7.26	3.02	8.25
⋮					
58	8.32	36.80	3.59	7.49	9.11
59	8.10	36.39	5.02	7.93	8.84
60	7.12	36.79	6.99	8.95	9.52

根据这 60 组数据用最小二乘法估得回归系数如下：

α \ b	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9
α_{300}	12.37	−0.35	−1.46	−2.11	0	0.089	0.074	0.050	0.193	0.110
α_{500}	17.43	−0.64	−1.55	−1.60	0.015	0.067	0.064	0.059	0.121	0.059

可以计算出 α_{300} 的回归残差平方和 $Q=12.12$，$\hat{\sigma}=0.49$；α_{500} 的回归

残差平方和 $Q=1.46$，$\hat{\sigma}=0.17$；这两个回归方程都是显著的，可以用来进行控制. 问题为要求

$$5.5\leqslant \alpha_{300}\leqslant 8.0$$

$$5.5\leqslant \alpha_{500}\leqslant 8.0$$

问 x_1, x_2, x_3 应在什么范围（这个范围称为瓷封区）？

从理论上看，这个范围是由几个二次曲面围成的，但由于曲面较难研究，故从实用角度看可利用切片法（即固定一个自变量，考察另两个自变量之间的曲线关系）找出瓷封区的一个典型部分即可. 根据以往的经验和节省钴的原则，选择钴在 5%到 6%之间，固定钴含量 5%、5.5%、6%，分别找出铜与镍的关系，最后得到钴、铜、镍应控制在下列范围内：

$$5\%\leqslant 钴\leqslant 6\%,\quad 3.5\%\leqslant 铜\leqslant 4.5\%$$

$$45\%\leqslant 铜+镍+钴\leqslant 46.5\%$$

这就是该课题组得到的结论. 在上述范围内取定成分，进行验证试验，制造出的合金材料的膨胀性能确实合乎要求，因而完成了研究任务.

（四）回归模型中的假设检验

现在来研究本节开头提出的问题(1)，即如何根据数据(5.1)判断(5.2)中随机项 e 的方差足够小使得所建立的回归方程(5.3)是有意义的. 这就是要检验下列假设

$$H_0: \beta_1=\beta_2=\cdots=\beta_p=0 \tag{5.12}$$

一旦假设(5.12)被拒绝，我们就认为所建立的回归方程(5.3)是有意义的或者说方程(5.3)是显著的. 这时就可利用(5.3)进行预测或控制.

有时还要判别某个自变量或某些自变量对 y 是否有显著影响，这个化为检验某个或某些回归系数为零. 即检验下列假设

$$H_0: \beta_{i_1}=\beta_{i_2}=\cdots=\beta_{i_s}=0 \tag{5.13}$$

（这里 s 是固定的，$1\leqslant s\leqslant p$，$1\leqslant i_1<\cdots<i_s\leqslant p$）.

假设(5.12)和(5.13)都是下列一般线性假设的特殊情形

$$H_0: H\beta=0, \tag{5.14}$$

这里 H 是 $s\times(p+1)$ 矩阵,各行线性无关.

我们先对假设(5.14)给出检验方法,然后分别用到假设(5.12)和(5.13)上去.

我们在"假定 B"下进行讨论,即我们假定

$$y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+e_i \quad (i=1,\cdots,n), \tag{5.15}$$

其中 $e_1,\cdots,e_n$ 相互独立同分布,共同分布是 $N(0,\sigma^2)$(σ 未知).

沿用前面的记号. 如 $X=(E,C)$.

记

$$W=\{\xi\in R^n:\xi=X\beta,\beta\in R^{p+1}\}$$

$$W_0=\{\xi\in R^n:\xi=X\beta,H\beta=0,\ \beta\in R^{p+1}\}$$

$$\hat{\xi}=\mathop{\mathrm{Proj}}_{W} Y,\quad \hat{\xi}_0=\mathop{\mathrm{Proj}}_{W_0} Y$$

则由定理 4.1 知统计量

$$F=\frac{\|\hat{\xi}-\hat{\xi}_0\|^2/(p+1-q)}{\|Y-\hat{\xi}\|^2/(n-p-1)} \tag{5.16}$$

在 H_0 下服从 $F(p+1-q,n-p-1)$ 分布,其中 q 为 W_0 的维数,易见 $q=p+1-s$,故 H_0 下

$$F=\frac{\|\hat{\xi}-\hat{\xi}_0\|^2/s}{\|Y-\hat{\xi}\|^2/(n-p-1)}\sim F(s,n-p-1)$$

下面我们来求 $\|Y-\hat{\xi}_0\|^2$ 与 $\|\hat{\xi}-\hat{\xi}_0\|^2$ 的计算公式. $\|Y-\hat{\xi}\|^2$ 就是最小二乘估计的残差平方和,即(见(5.7)式)

$$\|Y-\hat{\xi}\|^2=Q=l_{yy}-\sum_{i=1}^{p} l_{iy}\hat{\beta}_i$$

我们再来证明

$$\|\hat{\xi}-\hat{\xi}_0\|^2=\hat{\beta}'H'[H(X'X)^{-1}H']^{-1}H\hat{\beta}$$

事实上,若记 $\hat{\xi}_0=X\tilde{\beta}\in W_0$,则 $\tilde{\beta}$ 是约束 $H\beta=0$ 下的最小二乘估计,由定理 3.6 知有 $\lambda\in R^s$ 使

$$\begin{cases} X'X\tilde{\beta} - H'\lambda = X'Y \\ H\tilde{\beta} = 0 \end{cases}$$

但

$$X'X\hat{\beta} = X'Y$$

$$X'X(\tilde{\beta} - \hat{\beta}) = H'\lambda$$

$$\tilde{\beta} - \hat{\beta} = (X'X)^{-1}H'\lambda$$

两边都左乘 H，由 $H\tilde{\beta}=0$ 得

$$-H\hat{\beta} = H(X'X)^{-1}H'\lambda$$

$$\lambda = -[H(X'X)^{-1}H']^{-1}H\hat{\beta}$$

所以

$$\begin{aligned} \|\hat{\xi} - \hat{\xi}_0\|^2 &= \|X\hat{\beta} - X\tilde{\beta}\|^2 \\ &= (\tilde{\beta} - \hat{\beta})'X'X(\tilde{\beta} - \hat{\beta}) \\ &= (\tilde{\beta} - \hat{\beta})'H'\lambda \\ &= -\hat{\beta}'H'\lambda \\ &= \hat{\beta}'H'[H(X'X)^{-1}H']^{-1}H\hat{\beta} \end{aligned}$$

因此检验的统计量为

$$F = \frac{\hat{\beta}'H'[H(X'X)^{-1}H']^{-1}H\hat{\beta}/s}{\left(l_{yy} - \sum_{i=1}^{p} l_{iy}\hat{\beta}_i\right)/(n-p-1)} \qquad (5.17)$$

在 H_0 下 F 服从 $F(s, n-p-1)$ 分布，给定检验水平 α 后查相应自由度的 F 分布表，得上侧 α 分位点（即 $1-\alpha$ 分位数）c，可得检验 H_0 的否定域为

$$\{F > c\}$$

现在把上述一般方法用到假设(5.12)的检验上去，此时

$$H = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

这时直接可以看出

$$\|Y-\hat{\xi}_0\|^2 = \min_{\beta_0}\sum_{i=1}^{n}(y_i-\beta_0)^2$$

$$= \sum_{i=1}^{n}(y_i-\overline{y})^2 = l_{yy}$$

$$\|\xi-\hat{\xi}_0\|^2 = \|Y-\hat{\xi}_0\|^2 - \|Y-\hat{\xi}\|^2$$

$$= l_{yy} - \left(l_{yy} - \sum_{i=1}^{p} l_{iy}\hat{\beta}_i\right)$$

$$= \sum_{i=1}^{p} l_{iy}\hat{\beta}_i$$

记

$$U = \sum_{i=1}^{p} l_{iy}\hat{\beta}_i = l_{yy} - Q \tag{5.18}$$

称 U 为回归平方和，在 l_{yy} 中 U 占的比例大才说明各自变量 $x_1,\cdots,x_p$ 的确对 y 有显著影响．统计量

$$F=\frac{U/p}{Q/(n-p-1)} \tag{5.19}$$

在 H_0 下服从 $F(p,n-p-1)$ 分布，称 F 为回归的 F 统计量．查 F 分布上侧 α 分位点（即 $1-\alpha$ 分位数）c 可得检验 H_0 的否定域 $\{F>c\}$．

考察回归自变量对因变量的解释能力除可用(5.19)的回归 F 统计量外，应用中人们还常常用所谓“复相关系数”R，其定义为

$$R=\sqrt{\frac{U}{l_{yy}}} \tag{5.20}$$

这样

$$F=\frac{R^2/p}{(1-R^2)/(n-p-1)}$$

R^2 是回归平方和 U 在总平方和 l_{yy} 中占的比例大小．还可以证明

$$R^2 = \left[\frac{\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{y})}{\left\{\sum_{i=1}^{n}(y_i-\overline{y})^2\cdot\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2\right\}^{\frac{1}{2}}}\right]^2$$

即复相关系数是观测数据 $y_1, y_2, \cdots, y_n$ 与回归拟合值 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 之间的样本相关系数的绝对值. R 越大就说明回归方程越精确. 易见一元回归时 R 就是 $y_1, y_2, \cdots, y_n$ 与 $x_1, x_2, \cdots, x_n$ 之间的样本相关系数的绝对值.

现在来考虑假设(5.13). 先看一个特殊情形: $s=1$. 即单个系数为 0 的假设:

$$H_0: \beta_k = 0 \quad (k \text{ 是固定的}. 1 \leqslant k \leqslant p).$$

此时 $H=(0, \cdots, 0, 1, 0 \cdots 0)$ (1 是第 $k+1$ 个分量). $H(X'X)^{-1}H' = L^{-1}$ 的第 k 行第 k 列元素. 设 $L^{-1}=(l^{ij})_{p\times p}$, 则

$$\hat{\beta}'H'(H(X'X)^{-1}H')^{-1}H\hat{\beta} = (\hat{\beta}_k)^2 / l^{kk}$$

于是

$$F = \frac{(\hat{\beta}_k)^2}{l^{kk}Q/(n-p-1)} \tag{5.21}$$

在 H_0 下 F 服从 $F(1, n-p-1)$ 分布, 给定检验水平 α 后查出相应 F 分布的上侧 α 分位点(即 $1-\alpha$ 分位数) c 可得检验 H_0 的否定域为 $\{F>c\}$. 否定 H_0 则说明变量 x_k 对变量 y 有显著的影响. 这个假设也常用 t 分布来检验.

当假设(5.13)中的 $s>1$ 时使用统计量 F 的表达式(5.17)反而不方便. 直接使用(5.16)更好些. 这时

$$F = \frac{(Q_0 - Q)/s}{Q/(n-p-1)} \tag{5.22}$$

其中 $Q_0 = \|Y - \xi_0\|^2$, $Q = \|Y - \xi\|^2$, Q 是残差平方和, Q_0 也是一种"残差平方和", 后者是因变量 Y 对自变量 $\{x_1, \cdots, x_p\} - \{x_{i_1}, \cdots, x_{i_s}\}$ 的回归问题相应的残差平方和. 这样分别计算出 Q 和 Q_0 后 F 的值就很易得到了. 这样做可以避免计算矩阵 $X'X$ 的逆.

在假设(5.13)下, 统计量 $F \sim F(s, n-p+1)$. 给定检验水平 α. 设 c 是 $F(s, n-p+1)$ 的 $1-\alpha$ 分位数. 则假设(5.13)的否定域是 $\{F>c\}$.

例 5.3 表 5.3 中 y 为以 1948 年不变价格计算的英国进口

商品和劳务的指数，x_1 为以1948年价格计算的英国总产品的指数，x_2 为进口商品与普通英国产品价格指数的比．

计算得

$$n=9,\quad \sum y_i=1\,052,\quad \sum x_{i1}=1\,017,\quad \sum x_{i2}=954$$

$$\bar{y}=116.9,\quad \bar{x}_{\cdot 1}=113,\quad \bar{x}_{\cdot 2}=106$$

$$\sum y_i^2=124\,228,\quad \sum x_{i1}^2=115\,571,\quad \sum x_{i2}^2=101\,772$$

$$\sum y_i x_{i1}=119\,750,\quad \sum y_i x_{i2}=113\,433,\quad \sum x_{i1}x_{i2}=107\,690$$

表5.3 英国进口数据

年	y	x_1	x_2
1948	100	100	100
1949	106	104	99
1950	107	106	110
1951	120	111	126
1952	110	111	113
1953	116	115	103
1954	123	120	102
1955	133	124	103
1956	137	126	98

于是

$$L=\begin{pmatrix}650 & -112\\ -112 & 648\end{pmatrix},\quad L_{xy}=\begin{pmatrix}874\\ 1\,921\end{pmatrix}$$

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=L^{-1}L_{xy}=\begin{pmatrix}1.364\,23\\ 0.113\,88\end{pmatrix}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_{\cdot 1}-\hat{\beta}_2\bar{x}_{\cdot 2}=-49.329\,7$$

回归式为

$$\hat{y}=-49.329\,7+1.364\,2x_1+0.113\,9x_2$$

解释平方和(回归平方和)及残差平方和分别为

$$U=\hat{\beta}'L_{xy}=1\,183.34$$

$$Q=l_{yy}-U=\sum y_i^2-n\,\overline{y}^2-U=77.55$$

检验假设 $H_0:\beta_1=\beta_2=0$ 的统计量

$$F=\frac{U/2}{Q/(9-2-1)}=45.76$$

而 $F(2,6)$分布的 0.99 分位数(上侧 0.01 分位点)才为 10.925,所以回归方程是显著的.

检验假设 $H_0:\beta_1=0$ 的统计量

$$F=\frac{\hat{\beta}_1^2}{l^{11}Q/(9-2-1)}=\frac{1.364\,2^2}{(0.001\,586)(77.55)/6}=90.8$$

而(1,6)个自由度的 F 分布的 0.99 分位数(上侧 0.01 水平分位点)才为 13.7,所以 β_1 不为零,x_1 对 y 有显著影响.

检验假设 $H_0:\beta_2=0$ 的统计量

$$F=\frac{\hat{\beta}_2^2}{l^{22}Q/(9-2-1)}=\frac{0.113\,88^2}{(0.001\,59)(77.55)/6}=0.63$$

这个 F 值小于(1,6)自由度的 F 分布的 0.95 分位数(上侧 0.05 水平分位点)5.99,所以 F 不显著,不能拒绝 H_0,x_2 对 y 影响不大. 因此我们把变量 x_2 丢掉,建立 y 与 x_1 的新的回归方程:

$$\hat{y}=-35.09+1.345x_1$$

其 F 统计量为

$$F=\frac{\hat{\beta}_1 l_{1y}}{(l_{yy}-U)/(9-2)}=96.40$$

而 $F(1,7)$分布的 0.99 分位数(上侧 0.01 水平分位点)才为12.2,所以回归式仍是显著的.

例 5.4(生理节律模型) 为了测定一个人在 24 小时内的生理节律(例如血压(收缩压或舒张压)如何随时间而变化),一些学者提出了下列模型

$$f(t)=M+A\cos(\omega t+\phi),$$

其中 M 是基准值,A 是振幅,ϕ 是相位,ω 是角频率,例如 $\omega=360/24$. $f(t)$是所关心的生理指标. 问题是:设有观测值 $y_j=f(t_j)+\varepsilon_j$ $(j=1,2,\cdots,n)$,这里 t_j 是第 j 个观测时刻,$\varepsilon_1,\cdots,\varepsilon_n$ 是相互独立

同分布的随机项，$\varepsilon_j \sim N(0,\sigma^2)$（$\sigma$ 未知），如何估计 M,A,ϕ？（$0\leqslant\phi<360°$）.

解 易知

$$y_j = M + A\cos\phi \cdot \cos\omega t_j - A\sin\phi \cdot \sin\omega t_j + \varepsilon_j$$

故
$$y_j = M + \beta x_j + \gamma z_j + \varepsilon_j \quad (j=1,\cdots,n)$$

这里
$$x_j = \cos\omega t_j, \quad z_j = \sin\omega t_j,$$
$$\beta = A\cos\phi, \quad \gamma = -A\sin\phi \tag{5.23}$$

这便化成了二元线性回归模型.

我们可利用方程(5.6)求出 β,γ 的最小二乘估计 $\hat\beta,\hat\gamma$.

易知

$$l_{11} = \sum_1^n (x_j - \bar{x})^2, \quad l_{22} = \sum_1^n (z_j - \bar{z})^2,$$

$$l_{12} = \sum_{j=1}^n (x_j - \bar{x})(z_j - \bar{z}), \quad \left(\bar{x} = \frac{1}{n}\sum_1^n x_j, \bar{z} = \frac{1}{n}\sum_1^n z_j\right).$$

$$\left.\begin{aligned} l_{1y} &= \sum_1^n (x_j - \bar{x})(y_j - \bar{y}) \\ l_{2y} &= \sum_1^n (z_j - \bar{z})(y_j - \bar{y}) \end{aligned}\right. \quad \left(\bar{y} = \frac{1}{n}\sum_1^n y_j\right).$$

解方程(3.6)得

$$\hat\beta = \frac{l_{22}l_{1y} - l_{12}l_{2y}}{l_{11}l_{22} - l_{12}^2}, \quad \hat\gamma = \frac{-l_{12}l_{1y} + l_{11}l_{2y}}{l_{11}l_{22} - l_{12}^2}$$

$$\hat{M} = \bar{y} - \bar{x}\hat\beta - \bar{z}\hat\gamma.$$

从(5.23)可得到 A 和 ϕ 的估计 $\hat{A},\hat\phi$：

$$\hat{A} = \sqrt{(\hat\beta)^2 + (\hat\gamma)^2} \quad \left(\tan\hat\phi = -\frac{\hat\gamma}{\hat\beta}\right)$$

$$\hat\phi = \begin{cases} 360° - \theta, & \hat\beta > 0, \hat\gamma \geqslant 0 \\ \theta, & \hat\beta > 0, \hat\gamma < 0 \\ \theta + 180°, & \hat\beta \leqslant 0, \hat\gamma \geqslant 0 \\ 180° - \theta, & \hat\beta \leqslant 0, \hat\gamma < 0. \end{cases}$$

其中 $\theta=\arctan\left(\left|\frac{\hat{\gamma}}{\hat{\beta}}\right|\right)\quad(0\leqslant\theta\leqslant 90^{\circ})$

于是有非线性回归模型

$$\hat{y}=\hat{M}+\hat{A}\cos(\omega t+\hat{\phi})\tag{5.24}$$

这个方程是否有意义呢？要检验振幅 A 是否为 0. 这等价于检验 $H_0:\beta=\gamma=0$. 使用统计量

$$F=\frac{U/2}{Q/(n-3)}\quad(\text{参看}(5.19))$$

在 H_0 下 $F\sim F(2,n-3)$，给定 $\alpha\in(0,1)$，设 $F(2,n-3)$ 的 $1-\alpha$ 分位数是 λ，则 H_0 的否定域是 $\{F>\lambda\}$. 当 H_0 被拒绝时，所得到的回归方程(5.24)是有意义的. 在计算 F 时，注意 $U=l_{1y}\hat{\beta}+l_{2y}\hat{\gamma}$，$Q=l_{yy}-U$，这里 $l_{yy}=\sum_1^n(y_j-\bar{y})^2$.

在实际工作中，通常观测时刻是等间隔的，$t_j=\frac{j-1}{n}(j=1,2,\cdots,n)$ 且 $\omega=360^{\circ}$(最常见的情况是 $n=12$ 或 24)，这时上面的计算公式均大为简单. 实际上，$\sum_1^n x_j=\sum_{j=1}^n\cos\omega t_j=0$；$\sum_1^n z_j=\sum_1^n\sin\omega t_j=0$，$\sum_1^n x_jz_j=\sum_1^n(\cos\omega t_j)\sin\omega t_j=0$[①]，$\sum_1^n x_j^2=$

---

① 利用公式

$$\cos k\theta=\frac{\sin\left(k+\frac{1}{2}\right)\theta-\sin\left(k-\frac{1}{2}\right)\theta}{2\sin\frac{\theta}{2}}\quad(\text{当分母不是 }0)$$

知道

$$\sum_{k=0}^{n-1}\cos k\theta=\frac{2\sin\frac{n\theta}{2}\cos\frac{(n-1)\theta}{2}}{2\sin\frac{\theta}{2}}$$

故

$$\theta=\frac{360^{\circ}}{n}\text{ 时，}\sum_{k=0}^{n-1}\cos k\theta=0\quad(n\geqslant 2)$$

同理知

$$\sum_{k=0}^{n-1}\sin k\theta=0,\ \sum_{k=0}^{n-1}\cos k\theta\cdot\sin k\theta=0$$

$$\sum_{k=0}^{n-1}\left(\frac{1+\cos 2k\theta}{2}\right)=\frac{n}{2}\quad\left(\theta=\frac{360^\circ}{n}\right)$$

同理
$$\sum_1^n z_j^2=\frac{n}{2}$$

于是 $\hat{M}=\frac{1}{n}\sum_1^n y_j=\bar{y}$，$\hat{\beta}=\frac{1}{n}\sum_1^n x_jy_j$，$\hat{\gamma}=\frac{1}{n}\sum_1^n z_jy_j$，统计量 F 为

$$F=\frac{n(\hat{A})^2/2}{Q/(n-3)}$$

这里 $(\hat{A})^2=(\hat{\beta})^2+(\hat{\gamma})^2$，$Q=l_{yy}-n(\hat{A})^2$. 这些都是便于应用的简单公式。

（五）回归模型的残差分析

在讨论回归模型的假设检验（包括回归方程的显著性检验和回归系数的检验）时，我们有一个前提："假定 B"成立. 即假定模型(5.15)的随机项 $e_1,\cdots,e_n$ 相互独立同分布，共同分布是 $N(0,\sigma^2)$ (σ 未知). 怎样判别这个"假定 B"是否成立呢？

这是回归分析研究中很重要的问题. 现在介绍简单的处理办法.

根据数据(5.1)，我们可得回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$$

令
$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip}$$
$$\hat{e}_i=y_i-\hat{y}_i\quad(i=1,2,\cdots,n)$$

这些 $\hat{e}_i$ 叫做残差. 记 $\hat{e}=(\hat{e}_1,\cdots,\hat{e}_n)'$

则
$$\hat{e}=Y-X\hat{\beta}=Y-X(X'X)^{-1}X'Y=(I-P)Y,$$

这里 $P=X(X'X)^{-1}X'$，I 是单位阵. 设矩阵 P 的第 i 行第 i 列元素为 p_{ii}. 则在"假定 B"下 $\hat{e}$ 服从 n 维正态分布，且

$$\mathrm{Cov}(\hat{e},\hat{e})=\sigma^2(I-P)$$
$$E\hat{e}_i=0,\quad \mathrm{Var}(\hat{e}_i)=\sigma^2(1-p_{ii}).$$

令

$$\hat{\sigma}=\sqrt{\frac{Q}{n-p-1}}=\sqrt{\frac{1}{n-p-1}\sum_{1}^{n}(\hat{e}_i)^2}.$$

$$\gamma_i=\frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-p_{ii}}}\quad(i=1,\cdots,n)\tag{5.25}$$

这些 γ_i 叫做学生化残差. 数学上可以证明(参看韦博成等人的专著[15]),在“假定 B”下,只要 n 相当大,$\gamma_1,\cdots,\gamma_n$ 近似地相互独立同分布,且 γ_i 近似服从标准正态分布.

由此知道,如果“假定 B”成立,n 又比较大,则 $\gamma_1,\cdots,\gamma_n$ 中大约有$[0.95n]$个 γ_i 满足 $|\gamma_i|\leqslant 2$. 若根据实际数据(5.1)算出的 $\gamma_1,\cdots,\gamma_n$ 中有较多的 γ_i 不满足这个不等式,则应拒绝“假定 B”. 当 n 较大且不能拒绝“假定 B”时,一般应承认“假定 B”. 当 n 不太大时还应考察“残差图”,详细讨论见[15].

当“假定 B”被拒绝时,往下怎么办? 这时应对数据点$(x_{i1},\cdots,x_{ip},y_i)$进行过细的研究,要剔除“异常点”,这些属于回归诊断的内容,见[15].

(5.25)中的 p_{ii} 应如何计算呢? 可以证明下列公式成立:

$$p_{ii}=\frac{1}{n}+(x^{(i)}-\overline{x})'L^{-1}(x^{(i)}-\overline{x})$$

其中

$$L=(l_{ij})_{p\times p},\quad l_{ij}=\sum_{k=1}^{n}(x_{ki}-\overline{x}_{\cdot i})(x_{kj}-\overline{x}_{\cdot j})$$

$$\overline{x}=(\overline{x}_{\cdot 1},\cdots,\overline{x}_{\cdot p})',\quad x^{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})'$$

$$\overline{x}_{\cdot j}=\frac{1}{n}\sum_{k=1}^{n}x_{kj}\quad(1\leqslant j\leqslant p)$$

特别,$p=1$ 时

$$p_{ii}=\frac{1}{n}+\frac{1}{l_{xx}}(x_i-\overline{x})^2\quad(1\leqslant i\leqslant n)\tag{5.26}$$

其中

$$\bar{x}=\frac{1}{n}\sum_{1}^{n}x_i,\quad l_{xx}=\sum_{1}^{n}(x_i-\bar{x})^2$$

以上就是对回归模型进行残差分析，用以判别“假定 B”是否成立(当然，残差分析还有许多别的用处)．为了显示这种分析的必要性，我们提出下列例子．

例 5.5 设有下列两组数据：

第 1 组

x_i	1	2	3	4	5	6	7	8
y_i	2.0	0.5	1.0	3.5	2.0	4.5	5.0	3.5

第 2 组

x_i	1	2	3	4	5	6	7	8
z_i	0.5	1.0	2.5	2.0	5.5	3.0	3.5	4.0

今对这两组数据分别建立一元线性回归方程．经过简单的计算知 $l_{xx}=42$，$\sum_{1}^{8}y_i=\sum_{1}^{8}z_i=22$，$\sum_{1}^{8}y_i^2=\sum_{1}^{8}z_i^2=79$，$\sum_{1}^{8}x_iy_i=\sum_{1}^{8}x_iz_i=120$．由此知这两个组对应的回归方程是一样的，

$$\hat{y}=0.5+0.5x,\quad \hat{z}=0.5+0.5x$$

而且第 1 组对应的统计量 $F_1=6\dfrac{U_1}{Q_1}=7.875$($U_i$ 是第 i 组数据对应的回归平方和，Q_i 是第 i 组数据对应的残差平方和)，第 2 组对应的统计量 $F_2=6\dfrac{U_2}{Q_2}=7.875=F_1$．从而两个回归方程都在检验水平 $\alpha=0.05$ 上是“显著的”(F 分布相应的临界值是 $\lambda=5.99<7.875$)．这样看来，似乎同一个回归方程拟合两组数据都是合适的．其实不然．我们来看看各组数据的散点图(图 13 是第 1 组对应的散点图，图 14 是第 2 组对应的散点图)．

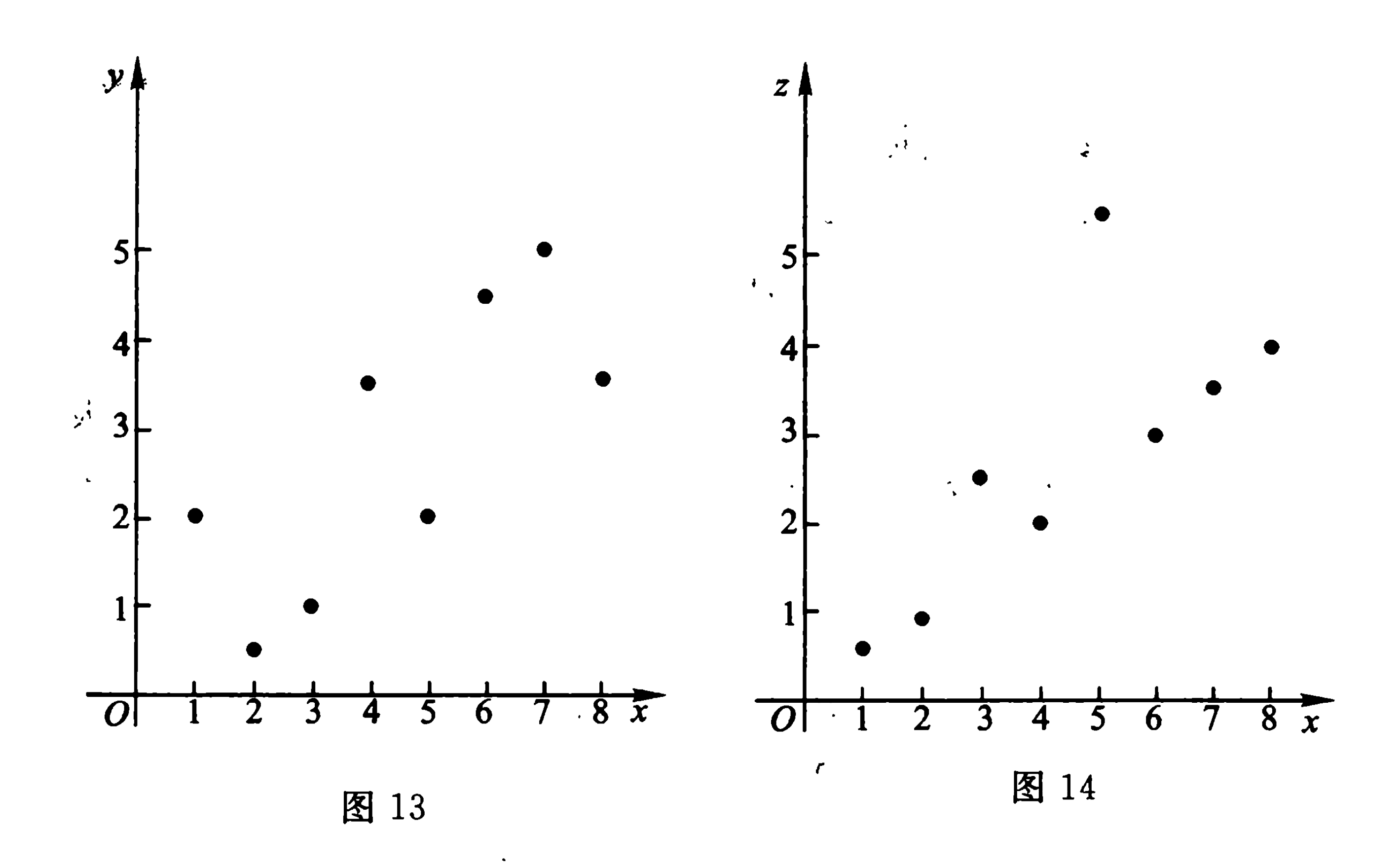


图 13　　图 14

从直观上不难看出，第 1 组的数据点基本上在一直线附近，第 2 组的数据点中有一个点(5，5.5)十分特别，很难说它和其他的点在一直线附近. 为确切起见，我们利用公式(5.25)和(5.26)计算各组数据对应的学生化残差，结果如下：

γ_i	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
第 1 组	1.134	−1.016	−0.956	0.929	−0.929	0.956	1.016	−1.134
第 2 组	−0.567	−0.508	0.478	−0.464	2.323	−0.478	−0.508	−0.567

从这张表看出，第 1 组对应的所有 γ_i 均满足 $|\gamma_i|<2$，故可认为"假定 B"成立，因而对这个组的回归方程的显著性检验是有效的($\alpha=0.05$). 对于第 2 组来说，有 γ_i 满足 $|\gamma_i|>2$，因而不能认为"假定 B"成立，对回归方程所做的显著性检验无效. 总之，用回归方程 $\hat{y}=0.5+0.5x$ 拟合第 1 组数据是有理由的，可用来进行预测或控制；用回归方程 $\hat{z}=0.5+0.5x$ 拟合第 2 组数据缺乏根据，不能用来进行预测或控制.

(六) 第二类回归简介

设 y 与 $x_1,x_2,\cdots,x_p$ 都是随机变量，要找 $f(x_1,x_2,\cdots,x_p)$ 使

$E[y-f(x_1,x_2,\cdots,x_p)]^2$ 达到最小，这就是第二类回归的问题.可以证明，只要 Ey^2 有限，取

$$\varphi(x_1,x_2,\cdots,x_p)=E(y|x_1,x_2,\cdots,x_p)$$

则对于一切 $f(x_1,x_2,\cdots,x_p)$ 有

$$E[y-\varphi(x_1,x_2,\cdots,x_p)]^2\leqslant E[y-f(x_1,x_2,\cdots,x_p)]^2$$

这个 φ 叫做回归函数，它一般不易求出，因此只好缩小函数类，限定 f 是线性函数，只要找 $\beta_0,\beta_1,\cdots,\beta_p$ 使得 $E[y-(\beta_0+\beta_1x_1+\cdots+\beta_px_p)]^2$ 达到最小.

记

$$\beta=\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_p\end{pmatrix},\quad X=\begin{pmatrix}x_1\\x_2\\\vdots\\x_p\end{pmatrix}$$

$$R_{xx}=\mathrm{Cov}(X,X),\quad R_{xy}=E[(X-EX)(y-Ey)]$$

则拟合误差的期望值

$$\begin{aligned}Q(\beta_0,\beta)&=E(y-\beta_0-\beta'X)^2\\&=E[(y-Ey)-\beta'(X-EX)+Ey-(\beta_0+\beta'EX)]^2\\&=E(y-Ey)^2+\beta'R_{xx}\beta-2\beta'R_{xy}+b^2\end{aligned}$$

其中 $b=Ey-(\beta_0+\beta'EX)$. $Q(\beta_0,\beta)$ 的最小值在 $\hat{\beta}_0,\hat{\beta}$ 处达到，$\hat{\beta}_0,\hat{\beta}$ 满足

$$R_{xx}\hat{\beta}=R_{xy}$$

$$\hat{\beta}_0=Ey-\hat{\beta}'EX$$

事实上，记 $\beta=\hat{\beta}+\delta$，则

$$\begin{aligned}Q(\beta_0,\beta)&\geqslant E(y-Ey)^2+\beta'R_{xx}\beta-2\beta'R_{xy}\\&=E(y-Ey)^2+(\hat{\beta}+\delta)'R_{xy}(\hat{\beta}+\delta)-2(\hat{\beta}+\delta)'R_{xy}\\&=E(y-Ey)^2+\hat{\beta}'R_{xx}\hat{\beta}-2\hat{\beta}'R_{xy}+\delta'R_{xx}\delta+\\&\quad 2\delta'(R_{xx}\hat{\beta}-R_{xy})\\&\geqslant E(y-Ey)^2+\hat{\beta}'R_{xx}\hat{\beta}-2\hat{\beta}'R_{xy}\\&=Q(\hat{\beta}_0,\hat{\beta})\end{aligned}$$

当 R_{xx} 非奇异时，

$$\hat{\beta}=R_{xx}^{-1}R_{xy}$$

$$\hat{\beta}_0=Ey-\beta'EX$$

这是理论上模型拟合的参数，但实际应用时只能得到如(5.1)那样的数据，理论上的 R_{xx} 与 R_{xy} 未知，只能估计. 令

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$\overline{x}_{\cdot i}=\frac{1}{n}\sum_{k=1}^{n}x_{ki},\quad i=1,2,\cdots,p$$

$$l_{ij}=\frac{1}{n}\sum_{k=1}^{n}(x_{ki}-\overline{x}_{\cdot i})(x_{kj}-\overline{x}_{\cdot j}),\quad i,j=1,2,\cdots,p$$

$$l_{iy}=\frac{1}{n}\sum_{k=1}^{n}(x_{ki}-\overline{x}_{\cdot i})(y_k-\overline{y}),\quad i=1,2,\cdots,p$$

$$L=(l_{ij})_{p\times p}$$

$$L_{xy}=(l_{iy})_{p\times 1}$$

用 L 和 L_{xy} 作为 R_{xx} 和 R_{xy} 的估计，并假定(5.1)的前 p 个列线性无关，则得 β_0,β 的估计为

$$\hat{\beta}=L^{-1}L_{xy}$$

$$\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}_{\cdot 1}-\cdots-\hat{\beta}_p\overline{x}_{\cdot p}$$

这与第一类回归的计算公式完全相同.

第二类回归中如果假定变量 $y,x_1,\cdots,x_p$ 的联合分布是多元正态分布，那么可以证明条件期望 $E(y|x_1,\cdots,x_p)$ 必为 $x_1,x_2,\cdots,x_p$ 的线性函数，因此这时“最优回归”必定是线性回归. 由于正态分布在实际工作中大量出现，所以线性回归在实际工作中有广泛的应用.

(七) 关于回归分析的几点注意事项

(1) 对实际问题，要根据问题的物理背景适当地选择模型. 在有些问题中已有成熟的理论或经验，如下面的例 5.4；如果没有理论的指导可以用不同的模型进行比较，一般地，可以比较模型的

回归 F 统计量的值，值越大，说明模型越合适；同时还可以看一看残差对自变量的散布图（对一元回归而言），图上的点应该随机地散布在某一条水平线周围，否则该模型还要加以改进.

例 5.6 放射性金（^{195}Au）对发炎的组织有亲和力，有时在诊断关节炎时把它用作示踪元素. 表 5.4 为注射 x 天后血液内的金元素残留百分比 y，共 $n=10$ 个血样.

表 5.4 示踪金元素残留量数据

注射天数 x	金残留的百分比 y
1	94.5
1	86.4
2	71
2	80.5
2	81.4
3	67.4
5	49.3
6	46.8
6	42.3
7	36.6

从散点图（见图 15）来看 x 与 y 的关系不太像是线性的，而且放射性元素的衰变从物理学来讲遵从负指数规律，所以拟合模型 $y=ae^{bx}$，即 $\ln y=\ln a+bx$，进行回归拟合得

$$\hat{y}=104.4e^{-0.1462x}$$

回归的 F 统计量值为 332.9，在 H_0 下 $P(F>332.9)<0.0001$. 如果直接拟合线性模型，有

$$\hat{y}=96.78-8.876x$$

其 F 统计量值为 178.8，仍在 0.0001 水平下是显著的，只比负指数模型略差. 这是因为曲线在局部可以用直线很好地逼近.

（2）使用回归方法一定要注意的是，即使建立了回归式并且统计检验证明相关关系成立，也只能说研究的变量是统计相关的，

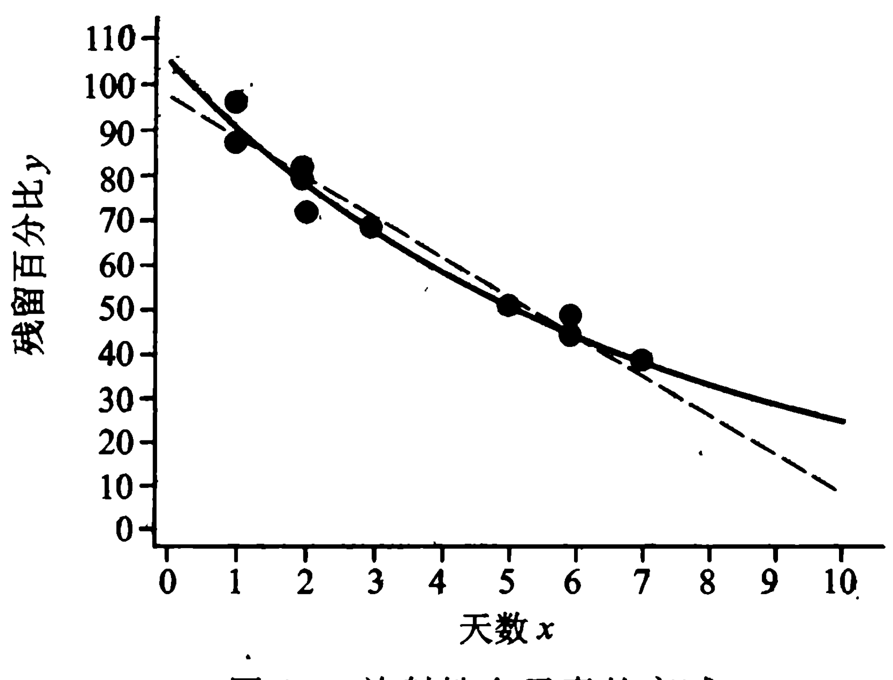


图 15　放射性金元素的衰减

而不能就此断定变量之间有因果关系，因为变量 y 与变量 x 线性相关有可能是它们都受另一个变量 z 的线性影响，而 y 和 x 之间可能根本没有什么因果关系. 比如国外有人发现喝咖啡的习惯与冠心病的发病高度统计相关，于是就有些文章声称咖啡是危及人们健康的一个隐患，后来进一步的研究发现，喝咖啡者较高的冠心病发病率是与吸烟和蔗糖摄入有关，而这两者与喝咖啡的习惯都高度相关.

(3) 回归的因变量选定后，自变量的选择十分重要. 如果遗漏了重要的解释变量，那么模型不能正确反映实际情况，而且参数估计是有偏的. 如果自变量之间有近似的线性相关，这时模型中的估计量很不稳定，参数估计的方差很大. 下一节讲的回归自变量选择方法只解决了部分问题，在选择自变量的问题上还是要进行深入的分析研究.

(4) 用回归方程作预报或控制，必须考虑对要研究的变量有影响的各种未引入模型中的因素是否有变化，即问题的背景是否发生变化. 如有变化，那么原来的经验公式在变化了的情况下可能不再适用，需要重新进行研究. 比如，政府政策的大的变动就可能影响到许多经济变量之间的关系.

*§6 回归自变量的选择

在线性回归分析中，常把所有可能影响因变量 y 的各种自变量 x_1，x_2，…，x_p 都列出来，根据对它们的观测数据建立 y 对 x_1，x_2，…，x_p 的回归方程. 这样做的一个缺点就是可能有一些自变量对 y 的影响很小甚至没有影响，但也列入了回归方程中，增加了取得数据和分析数据的负担，且得到的回归方程过于复杂，不便使用，另外，从随机误差项方差 σ^2 的估计 $\hat{\sigma}^2=Q/(n-p-1)$来看，多用自变量不一定导致较小的 $\hat{\sigma}^2$. 所以，我们应该建立这样的回归方程，它包含 x_1，x_2，…，x_p 中那些对 y 的影响不可忽视的变量，但不包含那些对 y 影响不大的变量，即从$\{x_1,x_2,\cdots,x_p\}$中选出一个适当的子集$\{x_{i_1},x_{i_2},\cdots,x_{i_l}\}$使得建立的回归方程恰含有这个子集的变量. 这就是著名的回归自变量的选择问题.

回归自变量的选择问题在应用上和理论上都十分重要，从上世纪 60 年代以来受到了广泛的注意. 这个问题的头一个困难是怎样比较不同选择的优劣，即最优选择的标准. 从不同的角度出发，可以有不同的比较标准，具体介绍如下.

以下将自变量 x_1，x_2，…，x_p 用代号 1，2，…，p 来表示，记 $I=\{1,2,\cdots,p\}$，I 是可取的自变量的全集. 设 A 是 I 的子集，则 A 对应一种自变量选择$\{x_i:i\in A\}$. 记$\#(A)$为 A 的元素个数. 若 $A=\{i_1,i_2,\cdots,i_l\}(i_1<i_2<\cdots<i_l)$，把 $X=(x_{ij})_{n\times p}$ 的第 i_1 列，第 i_2 列，…，第 i_l 列抽出来组成的矩阵记作 $X(A)$. 从数据$(X(A),Y)(Y=(y_1,y_2,\cdots,y_n)')$出发用最小二乘法可得回归方程

$$\hat{y}=\hat{\beta}_0(A)+\sum_{j\in A}\hat{\beta}_j(A)x_j \tag{6.1}$$

这叫做与 A 相应的回归方程，其残差平方和与回归平方和分别为

$$Q(A)=\|Y-X(A)\hat{\beta}(A)\|^2 \tag{6.2}$$

$$U(A)=l_{yy}-Q(A) \tag{6.3}$$

其中

$$\hat{\beta}(A)=(\hat{\beta}_0(A),\hat{\beta}_{i_1}(A),\cdots,\hat{\beta}_{i_l}(A))'$$

I 的子集共有 2^p 个，要从中找一个"最优"的，就要有一个比较子集优劣

的标准. 常见的比较标准有下列几种:

(1) 平均残差平方和最小. 即找 A 使

$$S^2=\frac{Q(A)}{n-1-\#(A)}$$

达到最小.

(2) 平均预测均方误差最小. 即找 A 使

$$J=\frac{n+\#(A)+1}{n-\#(A)-1}Q(A)$$

达到最小.

(3) C_p 统计量最小. 即找 A 使

$$C_p=\frac{Q(A)}{Q(I)/(n-p-1)}+2\cdot\#(A)-n$$

达到最小.

(4) AIC 准则. 即找 A 使

$$\mathrm{AIC}(A)=\ln Q(A)+2\cdot\frac{\#(A)}{n}$$

达到最小.

(5) BIC 准则. 即找 A 使

$$\mathrm{BIC}(A)=\ln Q(A)+\#(A)\cdot\frac{\ln n}{n}$$

达到最小.

这些准则都有一定的道理,很难说其中一个比另一个好,它们的理论基础仍值得探讨. 这样,我们要找最优回归方程就很难做到. 不过,实际工作者常常并不追求"最优",只要能找出比较优的回归方程就行了. 下面定义的"宜取的回归方程"就是一种比较优的回归方程.

给定 $A\subset I, i\in A, Q(A-\{i\})-Q(A)$ 可用来刻画变量 x_i 在回归式(6.1)中的作用(等价地可用 $U(A)-U(A-\{i\})$ 来刻画). 这个量叫做 x_i 关于 A 的偏回归平方和,记作 $P_i(A)$. 考虑

$$F_i(A)\overset{\mathrm{d}}{=}\frac{P_i(A)}{Q(A)/(n-1-\#(A))} \tag{6.4}$$

$F_i(A)$ 的大小反映了 x_i 在子集 A 中作用的大小. 给定两个数 $F_{进}, F_{出}, 0\leqslant F_{出}\leqslant F_{进}$.

定义 6.1 称子集 A 是可剔除的，如果存在 $i\in A$ 使得

$$F_i(A)\leqslant F_{出}$$

反之称 A 为不可剔除的．称子集 A 是排外的，若对任何 $i\in I-A$ 均有

$$F_i(A\cup\{i\})\leqslant F_{进}$$

称子集 A 是宜取的，若它是不可剔除的而且是排外的．与宜取子集 A 相应的回归方程(6.1)称为宜取的回归方程．宜取的回归方程是比较优的，大量实践表明，只要 $F_{进}$ 和 $F_{出}$ 取得适当(通常取 $F_{进}=F_{出}$ 在[3,8]内)，得到的宜取回归方程颇有实用价值，利用它进行预测和控制往往令人满意，因而受到实际工作者的欢迎．但是，宜取子集可能不止一个，究竟哪一个更好，宜取性和上面介绍的几种优良性准则有什么关系，这些问题都有待于进一步的理论探讨．

既然宜取的回归方程有实用价值，怎样从原始数据出发得到它呢？使用著名的逐步回归法就可得到一个宜取的回归方程．现在来介绍这个方法，其要点是从空集出发逐步引入自变量，同时剔除相对而言不重要的自变量，直到不能引入新的自变量也不能剔除为止．下面用数学归纳法给出逐步回归法的严格表述．用 A_t 代表第 t 步建立的回归方程所含的自变量的足标组成的集合．

第 0 步．令 $A_0=\varnothing$(空集)，这时回归方程是

$$\hat{y}=\bar{y} \tag{6.5}$$

第 1 步．考虑引入自变量．对所有 $i\in I$ 分别计算

$$F_i(\{i\})=\frac{Q(\varnothing)-Q(\{i_1\})}{Q(\{i\})/(n-2)}$$

(这里 $Q(\varnothing)=l_{yy}$)．找 i_1 满足

$$F_{i_1}(\{i_1\})=\max_{i\in I}F_i(\{i\})$$

如果 $F_{i_1}(\{i_1\})\leqslant F_{进}$，这时逐步回归过程结束，最后的回归方程是(6.5)；如果 $F_{i_1}(\{i_1\})>F_{进}$，这时引入自变量 x_{i_1}，令 $A_1=\{i_1\}$ 并建立与 A_1 相应的回归方程

$$\hat{y}=\hat{\beta}_0(A_1)+\hat{\beta}_{i_1}(A_1)x_{i_1} \tag{6.6}$$

第 2 步．考虑能否引入新的自变量．对所有 $i\in I-A_1$ 分别计算

$$F_i(A_1\cup\{i\})=\frac{Q(A_1)-Q(A_1\cup\{i\})}{Q(A_1\cup\{i\})/(n-3)}$$

找 i_2 使

$$F_{i_2}(A_1 \cup \{i_2\}) = \max_{i \in I - A_1} F_i(A_1 \cup \{i\})$$

若 $F_{i_2}(A_1 \cup \{i_2\}) \leqslant F_{进}$，则逐步回归过程结束，最后的回归方程是第 1 步得到的方程(6.6)；若 $F_{i_2}(A_1 \cup \{i_2\}) > F_{进}$，则引入变量 x_{i_2}，令 $A_2 = \{i_1, i_2\}$ 并建立与 A_2 相应的回归方程.

设上面的过程已进行了 t 步($t \geqslant 2$)，得到的自变量的足标集合为 A_t，相应的回归方程是

$$\hat{y} = \hat{\beta}_0(A_t) + \sum_{i \in A_t} \hat{\beta}_i(A_t) x_i \tag{6.7}$$

第 $t+1$ 步. 先考虑能否剔除. 对每个 $i \in A_t$ 计算

$$F_i(A_t) = \frac{Q(A_t - \{i\}) - Q(A_t)}{Q(A_t)/(n - 1 - \#(A_t))}$$

找 i_0 使

$$F_{i_0}(A_t) = \min_{i \in A_t} F_i(A_t)$$

如果 $F_{i_0}(A_t) \leqslant F_{出}$，则令 $A_{t+1} = A_t - \{i_0\}$，即剔除 x_{i_0}，建立与 A_{t+1} 相应的回归方程，然后转入第 $t+2$ 步；如果 $F_{i_0}(A_t) > F_{出}$，这时不剔除，考虑能否引入新变量. 对所有 $i \in I - A_t$ 计算

$$F_i(A_t \cup \{i\}) = \frac{Q(A_t) - Q(A_t \cup \{i\})}{Q(A_t \cup \{i\})/(n - 2 - \#(A))}$$

找 i_0 使

$$F_{i_0}(A_t \cup \{i_0\}) = \max_{i \in I - A_t} F_i(A_t \cup \{i\})$$

如果 $F_{i_0}(A_t \cup \{i_0\}) \leqslant F_{进}$，则逐步回归过程结束，最后的回归方程是(6.7)；如果 $F_{i_0}(A_t \cup \{i_0\}) > F_{进}$，则引入 x_{i_0}，令 $A_{t+1} = A_t \cup \{i_0\}$ 并建立与 A_{t+1} 相应的回归方程，转入第 $t+2$ 步.

这样我们用归纳法给出了逐步回归算法的全过程. 这个过程可直观地描述如下：自变量是逐步引入的. 每引入一个新变量，要先检查先前引入方程的变量是否要剔除，若有可剔除的，则从影响最小的开始剔除直到没有变量可剔除时再考虑能否引入新的变量，若能引入，则先引入影响最大的新变量，引入后马上考虑原先的自变量是否要剔除，……，到无变量可剔除时再考虑引入新的变量. 此过程进行到不能再引入新变量为止. 这时得到的回归方程便是宜取的回归方程. 算法的框图如图 16.

逐步回归既然是又引入又剔除，那么它就有一个能否在有限步停止的问题. 答案是肯定的，严格证明如下.

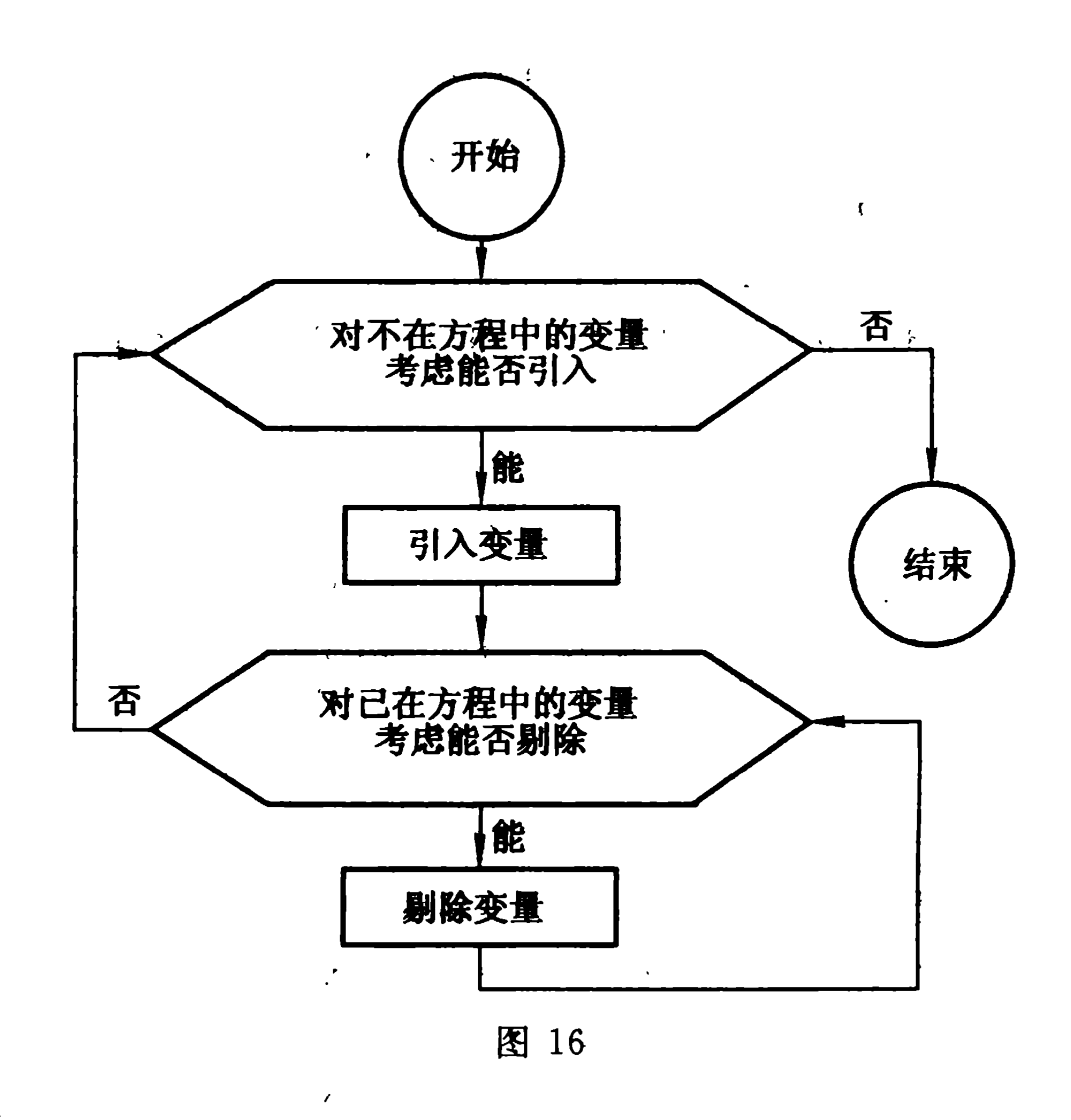


图 16

记

$$a_r=\frac{F_{出}}{n-r-1},\quad b_r=\frac{F_{进}}{n-r-1},\quad r=1,2,\cdots,p.$$

显然

$$0\leqslant a_r\leqslant b_r$$

且不难看出，子集 A 可剔除的充要条件是存在 $i\in A$ 使

$$\frac{Q(A-\{i\})-Q(A)}{Q(A)}=\frac{F_i(A)}{n-r-1}\leqslant a_r\quad (r=\#(A))\tag{6.8}$$

子集 A 不排外的充要条件是存在 $i\in I-A$ 使

$$\frac{Q(A)-Q(A\cup\{i\})}{Q(A\cup\{i\})}=\frac{F_i(A\cup\{i\})}{n-r-2}>b_{r+1}\quad (r=\#(A))\tag{6.9}$$

对 I 的子集 A，定义

$$\alpha(A)=\rho_0\rho_1\cdots\rho_{\#(A)}$$

其中 $\rho_0=1,\rho_i=1+b_i\quad (i=1,2,\cdots,p)$.

引理 6.1 设 $t\geqslant 0$.

(1) 若 A_t 可剔除，则

$$Q(A_t) \geqslant \frac{\alpha(A_{t+1})}{\alpha(A_t)} Q(A_{t+1})$$

(2) 若 A_t 不可剔除但不排外，则

$$Q(A_t) > \frac{\alpha(A_{t+1})}{\alpha(A_t)} Q(A_{t+1})$$

证明:(1) 若 A_t 可剔除，则第 $t+1$ 步剔除一个自变量得到 A_{t+1}，于是由(6.8)式可知

$$\frac{Q(A_{t+1}) - Q(A_t)}{Q(A_t)} \leqslant a_r \quad (r = \#(A_t))$$

从而

$$\begin{aligned} Q(A_t) &\geqslant \frac{1}{1+a_r} Q(A_{t+1}) \geqslant \frac{1}{1+b_r} Q(A_{t+1}) \\ &= \frac{1}{\rho_r} Q(A_{t+1}) \\ &= \frac{\alpha(A_{t+1})}{\alpha(A_t)} Q(A_{t+1}) \end{aligned}$$

(2) 若 A_t 不可剔除但不排外，则第 $t+1$ 步必引入一个新变量得到 A_{t+1}，于是由(6.9)式可知

$$\frac{Q(A_t) - Q(A_{t+1})}{Q(A_{t+1})} > b_{r+1} \quad (r = \#(A_t))$$

从而

$$Q(A_t) > (1 + b_{r+1}) Q(A_{t+1}) = \frac{\alpha(A_{t+1})}{\alpha(A_t)} Q(A_{t+1})$$

证毕.

引理 6.2 若存在 k 满足 $0 \leqslant t \leqslant k < l$ 且 A_k 是不可剔除的，A_{l-1} 不是宜取的，则

$$Q(A_t) > \frac{\alpha(A_l)}{\alpha(A_t)} Q(A_l)$$

证明:多次使用引理 6.1 中的不等式即得.

引理 6.3 设 $0 \leqslant t < l$, A_{l-1} 不是宜取的且 $\#(A_t) \leqslant \#(A_l)$，则 $Q(A_t) > Q(A_l)$.

证明:从条件 $\#(A_t) \leqslant \#(A_l)$ 及 $t < l$ 知必存在 k，$t \leqslant k < l$，使得 A_k 是不可剔除的. 由引理 6.2 即知

$$Q(A_t) > \frac{\alpha(A_l)}{\alpha(A_t)} Q(A_l) \geqslant Q(A_l)$$

证毕.

定理 6.1 若临界值 $F_{进} \geqslant F_{出} \geqslant 0$,则逐步回归法的筛选变量过程必在有限步后停止.

证明①:用反证法. 假若筛选过程永无止境,则得到子集列 $A_0, A_1, A_2, \cdots$. 由于不同的子集至多有 2^p 个,故存在 $t<l$ 使得 $A_t = A_l$,从而 $Q(A_t) = Q(A_l)$,这与引理 6.3 的结论相矛盾,故逐步回归法的变量筛选过程必在有限步后停止.

证毕.

逐步回归的计算有简单的递推公式,所以计算量并不大,许多统计软件包中都有计算逐步回归的程序,实际工作者很乐于采用.

§7 逻辑斯谛(logistic)回归模型

在前面介绍的经典线性回归模型里,因变量(响应变量)是连续变量. 在实际工作中(特别是社会现象的研究)常遇到因变量只取分类值,尤其是只取二分类值(即 0 或 1)的情形,这时就不能用前面的处理方法了. 例如,用 x 表示一个家庭的年收入,$Y=1$ 表示该家庭在一段时间内购买某种耐用消费品(例如汽车),$Y=0$ 表示不购买这种耐用消费品. 我们要研究的是概率 $P(Y=1)$ 与 x 的关系.

更一般地,若随机变量 Y 只取值 0 或 1,有若干个变量 $x_1, x_2, \cdots, x_k$ 影响 Y 的取值,我们关心的是概率 $p=P(Y=1)$ 是如何依赖于 $x_1, x_2, \cdots, x_k$?

对 p 的研究等价于对 $\frac{p}{1-p}$ 的研究,因为后者是 p 的严格增连续函数. $\frac{p}{1-p}$ 叫作发生比或优比(odds ratio).

定义 7.1 如果有下列关系式:

$$\ln \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{7.1}$$

---

① 这里给出的证明是根据江泽培先生的证法改写的.

（其中 $\beta_0,\beta_1,\cdots,\beta_k$ 是常数），则称二分类值变量 $Y(p=P(Y=1))$ 与自变量 $x_1,\cdots,x_k$ 的关系符合逻辑斯谛模型.

为了体现概率 p 与 $x_1,\cdots,x_k$ 的联系，常写 $p=P(Y=1\mid x_1,\cdots,x_k)$.(7.1)有下列等价形式：

$$P(Y=1\mid x_1,\cdots,x_k)=\frac{\exp\left\{\beta_0+\sum_1^k\beta_i x_i\right\}}{1+\exp\left\{\beta_0+\sum_1^k\beta_i x_i\right\}}$$

在(7.1)中，$\beta_0,\beta_1,\cdots,\beta_k$ 通常是未知的，需要利用数据进行估计. 一旦这些参数的值确定了，(7.1)式就可用来对 p 进行预测，也可用来对各自变量的重要性进行评价.

为简单计，以下只考虑 $k=1$（即一个自变量）的情形，用 x 表示 x_1，这时(7.1)化为

$$\ln\frac{p}{1-p}=\beta_0+\beta_1 x \tag{7.2}$$

令 $p(x)=P(Y=1\mid x)$，则

$$p(x)=\frac{\exp\{\beta_0+\beta_1 x\}}{1+\exp\{\beta_0+\beta_1 x\}} \tag{7.3}$$

怎样估计未知参数 β_0,β_1 呢？通常有两个办法：最大似然估计法和加权最小二乘法.

最大似然法：设有下列数据：$x=x_i$ 时 Y 的值是 $y_i(i=1,2,\cdots,n)$，$y_i=0$ 或 1. 应注意，这里 x_i 是自变量 x 的第 i 个值，不是(7.1)中的第 i 个自变量！

显然，

$$P(Y=y_i\mid x_i)=(p(x_i))^{y_i}(1-p(x_i))^{1-y_i}$$

于是观测值 $(x_1,y_1),\cdots,(x_n,y_n)$ 对应的似然函数是

$$L(\beta_0,\beta_1)=\prod_{i=1}^n(p(x_i))^{y_i}(1-p(x_i))^{1-y_i}$$

从而

$$\ln L(\beta_0,\beta_1)=\sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)-\sum_{i=1}^{n}\ln(1+e^{\beta_0+\beta_1 x_i})$$

令

$$\frac{\partial \ln L(\beta_0,\beta_1)}{\partial \beta_i}=0 \quad (i=0,1)$$

得似然方程组:

$$\sum_{i=1}^{n}\left(y_i-\frac{\exp\{\beta_0+\beta_1 x_i\}}{1+\exp\{\beta_0+\beta_1 x_i\}}\right)=0$$

$$\sum_{i=1}^{n}\left(y_i-\frac{\exp\{\beta_0+\beta_1 x_i\}}{1+\exp\{\beta_0+\beta_1 x_i\}}\right)x_i=0$$

若$(\hat{\beta}_0,\hat{\beta}_1)$是似然方程组的根,且 $x_1,\cdots,x_n$ 不全相等,则似然方程组的根是唯一的,而且$(\hat{\beta}_0,\hat{\beta}_1)$是 $L(\beta_0,\beta_1)$的最大值点,因而 $\hat{\beta}_0,\hat{\beta}_1$ 分别是 β_0,β_1 的最大似然估计(可以证明:$\ln L(\beta_0,\beta_1)$是二元严格凹函数). 但应注意的是,似然方程组有时无根(例如所有 y_i 都是 1 的情形). 在 SAS 和 SPSS 等国际流行的软件包里都有计算最大似然估计 $\hat{\beta}_0,\hat{\beta}_1$ 的程序.

加权最小二乘法:此法对数据有些特殊要求. 设 $x=x_i$ 时对 Y 作了 n_i 次观测(n_i 较大),其中事件$\{Y=1\}$发生了 r_i 次($i=1,2,\cdots,m$),这里 $x_1,\cdots,x_m$ 两两不同. 通常用

$$z_i=\ln\frac{r_i+0.5}{n_i-r_i+0.5} \tag{7.4}$$

作为 $\ln\dfrac{p(x_i)}{1-p(x_i)}$的估计值($i=1,2,\cdots,m$).

令

$$v_i=\frac{(n_i+1)(n_i+2)}{n_i(r_i+1)(n_i-r_i+1)} \quad (i=1,\cdots,m) \tag{7.5}$$

$$\widetilde{Q}(\beta_0,\beta_1)=\sum_{i=1}^{m}\frac{1}{v_i}(z_i-\beta_0-\beta_1 x_i)^2$$

使$\widetilde{Q}(\beta_0,\beta_1)$达到最小值的$\widetilde{\beta}_0,\widetilde{\beta}_1$ 分别称为β_0,β_1 的加权最小二乘估计. 这里$\dfrac{1}{v_1},\cdots,\dfrac{1}{v_m}$就是所谓的权. 可以证明,加权最小二乘估计存

在且唯一. 令$\frac{\partial\tilde{Q}(\beta_0,\beta_1)}{\partial\beta_i}=0(i=0,1)$,得方程组

$$\beta_0\sum_1^m\frac{1}{v_i}+\beta_1\sum_1^m\frac{x_i}{v_i}=\sum_1^m\frac{z_i}{v_i}$$

$$\beta_0\sum_1^m\frac{x_i}{v_i}+\beta_1\sum_1^m\frac{x_i^2}{v_i}=\sum_1^m\frac{x_iz_i}{v_i}$$

解此方程组,可得加权最小二乘估计如下:

$$\hat{\beta}_0=\frac{1}{l_1l_3-(l_2)^2}(l_5l_3-l_2l_4) \tag{7.6}$$

$$\hat{\beta}_1=\frac{1}{l_1l_3-(l_2)^2}(l_1l_4-l_2l_5) \tag{7.7}$$

这里

$$l_1=\sum_1^m\frac{1}{v_i},\quad l_2=\sum_1^m\frac{x_i}{v_i},\quad l_3=\sum_1^m\frac{x_i^2}{v_i},$$

$$l_4=\sum_1^m\frac{x_iz_i}{v_i},\quad l_5=\sum_1^m\frac{z_i}{v_i}$$

加权最小二乘估计是基于什么思想导出的呢? 本来应用$\frac{r_i}{n_i-r_i}$作为$\frac{p(x_i)}{1-p(x_i)}$的估计,为了避免分子和分母出现零,所以用$\frac{r_i+0.5}{n_i-r_i+0.5}$作为$\frac{p(x_i)}{1-p(x_i)}$的估计. 可以证明(基于概率论中的极限定理), $z_i=\ln\frac{r_i+0.5}{n_i-r_i+0.5}$ 近似服从正态分布 $N\left(\ln\frac{p(x_i)}{1-p(x_i)},\frac{1}{n_ip(x_i)(1-p(x_i))}\right)$.

于是 $z_i=\ln\frac{p(x_i)}{1-p(x_i)}+\varepsilon_i$,这里 ε_i 近似服从 $N(0,\Delta_i)$, $\Delta_i=(n_ip(x_i)(1-p(x_i)))^{-1}$、自然想到用 v_i 估计 Δ_i(v_i 的定义见(7.5)). 利用(7.2)得

$$z_i=\beta_0+\beta_1x_i+\varepsilon_i\quad(i=1,\cdots,m).$$

注意 $v_1,\cdots,v_m$ 不一定相等. 令 $\tilde{\varepsilon}_i=v_i^{-\frac{1}{2}}\varepsilon_i\quad(i=1,\cdots,m)$. 则

$$\frac{1}{\sqrt{v_i}}z_i=\frac{1}{\sqrt{v_i}}(\beta_0+\beta_1 x_i)+\tilde{\varepsilon}_i$$

这里 $\tilde{\varepsilon}_1,\cdots,\tilde{\varepsilon}_m$ 的方差近似相等. 仿效通常的最小二乘法的想法，应找 β_0,β_1 使得平方和 $\sum_1^m\left[\frac{1}{\sqrt{v_i}}z_i-\frac{1}{\sqrt{v_i}}(\beta_0+\beta_1 x_i)\right]^2$ 达到最小，这个平方和就是上文的 $\tilde{Q}(\beta_0,\beta_1)$. 因而使用加权最小二乘估计是有道理的.

例 7.1(社会调查) 一个人在家是否害怕生人来? 我们研究人的文化程度对此问题的影响. 因变量

$$Y=\begin{cases}1, & 害怕\\ 0, & 不害怕\end{cases}$$

自变量 x 是文化程度，取 4 个可能的值：x_1,x_2,x_3,x_4，这里

$x_1=0$ 表示文盲，$x_2=1$ 表示小学文化程度，$x_3=2$ 表示中学文化程度，$x_4=3$ 表示大专以上文化程度.

根据某地的一项社会调查报告，有下列数据：

自变量(x)	不害怕($Y=0$)人数	害怕($Y=1$)人数
0	11	7
1	45	32
2	664	422
3	168	72

我们可用逻辑斯谛(logistic)回归模型对上述数据进行统计分析. 用 $p(x)$ 表示一个人文化程度是 x 时害怕生人的概率，即 $p(x)=P(Y=1|x)$. 考虑模型

$$\ln\frac{p(x)}{1-p(x)}=\beta_0+\beta_1 x,$$

我们用加权最小二乘法估计 β_0,β_1.

根据上列数据，利用(7.4)和(7.5)可算出 $z_1=-0.3847$，$z_2=-0.3269$，$z_3=-0.4515$，$z_4=-0.8425$，$v_1=0.2199$，$v_2=$

0.052 7，v_3＝0.003 87，v_4＝0.019 7.

利用(7.6)和(7.7)可得到 $\hat{\beta}_0=0.013$，$\hat{\beta}_1=-0.25$，于是有回归方程

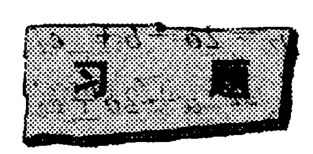
$$\ln\frac{p(x)}{1-p(x)}=0.013-0.25x$$

即

$$P(Y=1|x)=\frac{\exp\{0.013-0.25x\}}{1+\exp\{0.013-0.25x\}}$$

由此知，一个人文化程度越高，害怕生人的概率越低.

以上只对一个自变量的逻辑斯谛回归作了初步介绍. 若要了解更多的知识，请看张尧庭编著的《定性资料的统计分析》一书（广西师范大学出版社，1991).



1. 证明定理 2.1 中的最小二乘估计量 $\hat{b}$ 还可以写成以下三种形式：

(1) $$\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}$$

(2) $$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

(3) $$\hat{b}=\frac{\sum_{i=1}^{n}x_i(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

2. 设 $\hat{a}$，$\hat{b}$ 为一元线性回归系数 a，b 的最小二乘估计，记 $\hat{y}_i=\hat{a}+\hat{b}x_i$，$i=1,2,\cdots,n$，证明

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=0$$

$$\sum_{i=1}^{n}x_i(y_i-\hat{y}_i)=0$$

3. 设对同一组点(x_i,y_i)，$i=1,2,\cdots,n$ 分别用最小二乘法拟合一次式 $y=a+bx$和二次式 $y=\beta_0+\beta_1 x+\beta_2 x^2$，说明为什么二次式拟合的残差平方和不会大于一次式拟合的残差平方和.

4. 为了考察维尼伦纤维的耐热水性能，需要探讨甲醛浓度(g/L)对指标"缩醛化度"(克分子%)的影响. 安排了一批试验，得数据如下：

甲醛浓度 x	18	20	22	24	26	28	30
缩醛化度 y	26.86	28.35	28.75	28.87	29.75	30.00	30.36

从经验和理论知二者的关系是近似线性的，即有

$$y=\beta_0+\beta_1 x+e$$

试找出 β_0,β_1 的最小二乘估计，并画出原始数据及回归直线的图形.

5. 设

$$y_1=a+e_1$$
$$y_2=2a-b+e_2$$
$$y_3=a+2b+e_3$$

其中 a,b 是未知参数，e_1,e_2,e_3 相互独立且 $Ee_i=0$，$\mathrm{Var}(e_i)=\sigma^2$ $(i=1,2,3)$. 试求出 a,b 的最小二乘估计.

6. 设

$$y_i=\beta_0+\beta_1 x_i+\beta_2(3x_i^2-2)+e_i \quad (i=1,2,3)$$

其中 x_1,x_2,x_3 已知，β_0,β_1,β_2 是未知参数，y_1,y_2,y_3 是观测值，e_1,e_2,e_3 两两不相关，$Ee_i=0$，$\mathrm{Var}(e_i)=\sigma^2$ $(i=1,2,3)$. 试求出 β_0,β_1,β_2 的最小二乘估计.

7. 某医院用光电比色计检验尿汞时，得到的尿汞含量(mg/L)与消光系数读数的结果如下：

尿汞含量 x	2	4	6	8	10
消光系数 y	64	138	205	285	360

假定 y 与 x 有近似线性关系

$$y=\beta_0+\beta_1 x+e$$

各次测量是独立的且 $e\sim N(0,\sigma^2)$. 试求出 β_0,β_1 的最小二乘估计并检验 β_1 是否为零(检验水平取 0.05).

8. 下表中列出了 x 和 y 的两个样本的均值、标准差和相关系数：

样本	样本量 n	$\overline{x}$	$\overline{y}$	S_x	S_y	r_{xy}

1	600	5	12	2	3	0.6
2	400	7	10	3	4	0.7

试计算两个样本合在一起后 x 与 y 之间的样本相关系数，并说明这个相关系数小于原来任何一个样本的相关系数.

9. 下表是 20 世纪 60 年代美国洲际弹道导弹数的增长概况：

年 x	60	61	62	63	64	65	66	67	68	69
导弹数 y	18	63	294	424	834	854	904	1 054	1 054	1054

用逻辑斯谛模型 $y=L/(1+e^{a+bx})$ 拟合这些数据，取 $L=1\,060$. 画出原始数据与拟合曲线的图形.

10. 设随机向量 $Y\sim N(X\beta,\sigma^2 I_n)$，其中 X 为已知的 $n\times p$ 矩阵($n\geqslant p$)，试写出 Y 的密度函数并求出得到 Y 的观测值后参数 β 与 σ^2 的最大似然估计.

11. 设 X 是 $n\times p$ 矩阵，$\mathrm{Rank}(X)=p$($\mathrm{Rank}(X)$ 表示 X 的秩)，$P_X\overset{\mathrm{d}}{=}X(X'X)^{-1}X'$，又 $\tilde{X}$ 是 X 的一些列组成的 $n\times q$ 矩阵($q\leqslant p$)，$P_{\tilde{X}}\overset{\mathrm{d}}{=}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$，试证明

$$P_X\geqslant P_{\tilde{X}}$$

(即 $P_X-P_{\tilde{X}}$ 非负定).

12. 设 X 是 $n\times p$ 矩阵，$\mathrm{Rank}(X)=p$，又 $P_X=X(X'X)^{-1}X'$，试证明下列结论：

(1) n 维列向量 y 在子空间 $\mu(X)$ 上的投影 $\xi=P_X y$.

(2) 记 $P_X=(h_{ij})_{n\times n}$，则

$$0\leqslant h_{ii}\leqslant 1\quad (i=1,2,\cdots,n)$$

(3) 如果 X 的第一列元素全是 1，则

$$h_{ii}\geqslant\frac{1}{n}\quad (i=1,2,\cdots,n)$$

13. 设 X 是 $n\times p$ 矩阵，试证明以下结论：

(1) $\mathrm{Rank}(X'X)=\mathrm{Rank}(X')$

(2) $\mu(X'X)=\mu(X')$

(3) $(X'X)(X'X)^{-}X'=X'$，其中 $(X'X)^{-}$ 是 $X'X$ 的广义逆.

(4) n 维列向量 y 在子空间 $\mu(X)$ 的投影 $\xi=X(X'X)^{-}X'y$.

(5) 为了 β 是方程

$$X'X\beta = X'y$$

(y 为 n 维列向量)的解,必须且只须

$$\beta = (X'X)^{-}X'y + (I_p - (X'X)^{-}(X'X))u$$

其中 u 为任意的 p 维列向量.

14. 考虑约束线性模型

$$\begin{cases} Y = \theta + e \\ A\theta = 0 \end{cases}$$

其中 A 为 $r \times n$ 矩阵($n > r$),秩为 r. 证明 θ 的约束最小二乘估计为

$$\hat{\theta} = [I - A'(AA')^{-1}A]Y$$

15. 考虑约束线性模型

$$\begin{cases} Y = X\beta + e \\ H\beta = r \end{cases}$$

其中 X 为满秩 $n \times p$ 矩阵($n > p$),H 为满秩 $q \times p$ 矩阵($q < p$),则 β 的约束最小二乘估计有表达式

$$\hat{\beta} = \tilde{\beta} + (X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}(r - H\tilde{\beta})$$

其中 $\tilde{\beta} = (X'X)^{-1}X'Y$ 为无约束的普通最小二乘估计.

16. 在药物的临床研究中,病人对新药 B 的反应 y 与其对标准药物 A 的反应 x_1 及病人的心率 x_2 有关,下表为观测到的 10 组数据:

x_1	1.9	0.8	1.1	0.1	−0.1	4.4	4.6	1.6	5.5	3.4
x_2	66	62	64	61	63	70	68	62	68	66
y	0.7	−1.0	−0.2	−1.2	−0.1	3.4	0.0	0.8	3.7	2.0

拟合 y 对 x_1, x_2 的线性回归式 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ 并估计误差方差 σ^2.

17. 设

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \cdots, n$$

其中 $e_1, e_2, \cdots, e_n$ 独立同 $N(0, \sigma^2)$ 分布(σ^2 未知),设 $\hat{\beta}_0, \hat{\beta}_1$ 为 β_0, β_1 的最小二乘估计,试证明

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

并推导检验 $H_0:\beta_0=0$ 用的统计量并说明其分布.

18. 对第 7 题检验 $H_0:\beta_0=0$(水平 0.05).

19. 许多人认为人的寿命与手掌上的生命线长度有线性关系,下面的 50 对观测值是否支持了这种想法?其中 x 为寿命(年),y 为左手掌上生命线的长度(cm).

x	19	40	42	42	47	49	50	54	56	56
y	9.75	9.00	9.60	9.75	11.25	9.45	11.25	9.00	7.95	12.00
x	57	57	58	61	62	62	65	65	65	66
y	8.10	10.20	8.55	7.20	7.95	8.85	8.25	8.85	9.75	8.85
x	66	66	67	68	68	68	69	69	70	71
y	9.15	10.20	9.15	7.95	8.85	9.00	7.80	10.05	10.50	9.15
x	71	71	72	73	74	74	75	75	75	76
y	9.45	9.45	9.45	8.10	8.85	9.60	6.45	9.75	10.20	6.00
x	77	80	82	82	82	83	86	88	88	94
y	8.85	9.00	9.75	10.65	13.20	7.95	7.95	9.15	9.75	9.00

统计量 $\sum x=3\ 333$, $\sum x^2=231\ 933$, $\sum xy=30\ 549.75$, $\sum y=459.9$, $\sum y^2=4\ 308.57$.

20. 某合金材料含有两种主要成分,为了探讨这两种成分含量之和 x 与合金的膨胀系数 y 之间的关系,做了一批试验得数据如下:

x	37.0	37.5	38.0	38.5	39.0	39.5	40.0	40.5
y	3.40	3.00	3.00	3.27	2.10	1.83	1.53	1.70
x	41.0	41.5	42.0	42.5	43.0			
y	1.80	1.90	2.35	2.54	2.90			

画出数据的散点图,问 y 与 x 之间是否有相关关系,如有,则求出经验公式.

21. 炼钢基本上是个氧化脱碳的过程,钢液原来的含碳量的多少直接影响到冶炼时间的长短. 下表是某种平炉 34 炉的熔毕碳(即全部炉料熔化完毕时钢液的含碳量)与精炼时间(从熔毕到出钢所需的时间). 用 x 表示熔毕碳(单位:0.01%),y 表示精炼时间(分). 试问:y 与 x 是否有线性相关关系?若有,请找出回归方程. 若测得某炉熔毕碳为 145(即 1.45%),试估计该炉所需的精炼时间(置信度 95%). 如果想控制炼钢时间在 240 分以下,95%置信度下熔毕碳不能高于多少?

编号	熔毕碳 x	精炼时间 y	编号	熔毕碳 x	精炼时间 y
1	180	200	18	116	100
2	104	100	19	123	110
3	134	135	20	151	180
4	141	125	21	110	130
5	204	235	22	108	110
6	150	170	23	158	130
7	121	125	24	107	115
8	151	135	25	180	240
9	147	155	26	127	135
10	145	165	27	115	120
11	141	135	28	191	205
12	144	160	29	190	220
13	190	190	30	153	145
14	190	210	31	155	160
15	161	145	32	177	185
16	165	195	33	177	205
17	154	150	34	143	160

22. 回归分析常用在两个量 y,x 密切相关而其中一个(比如 y)难以测量时,用测量另一变量(x)来估计此变量(y)的值. 比如,测量不规则物体的体积可能十分困难,而称其重量却很容易. 下表为 5 到 8 岁的 18 个小孩的体重 x(kg)与体积 y(dm^3)的数据:

x	17.1	10.5	13.8	15.7	11.9	10.4	15.0	16.0	17.8
y	16.7	10.4	13.5	15.7	11.6	10.2	14.5	15.8	17.6
x	15.8	15.1	12.1	18.4	17.1	16.7	16.5	15.1	15.1
y	15.2	14.8	11.9	18.3	16.7	16.6	15.9	15.1	14.5

(1) 画出数据的散点图并拟合回归直线 $y=\hat{a}+\hat{b}x$.

(2) 求 $E(y|14.0)$的 95%置信区间.

(3) 求重 14.0 kg 的小孩的体积的 95%预测区间.

23. 某公司为了推销商品,研究广告费用 x 与获得的纯利润 y 之间的关系,以确定最佳的广告策略. 调查以往的情况,有以下数据:

x	1	1	2	2	2	3
y	14.80	15.90	20.20	20.00	18.55	22.20
x	3	4	4	4	5	5
y	20.90	21.00	18.30	20.70	16.10	14.75

(1) 画出数据的散点图,试找出 y 对 x 的适当的相关关系式(研究 y 对 x 的一次式、二次式、三次式并比较其优劣).

(2) 最好的广告策略是什么?

第五章 试验设计与方差分析

§1 全面试验的方差分析

试验设计和方差分析是数理统计学的两个重要组成部分.

在工农业生产和科学试验(或叫实验)中常常要通过试验研究若干个因素对一个(或多个)指标的影响.

例 1.1 烟灰砖试验.

某建筑材料研究所为了贯彻"综合利用"的精神,用工业废料——烟灰来做砖.想通过试验摸索好的生产工艺以保证烟灰砖的质量合乎要求.质量指标是折断力,它可用每平方厘米多少千克来表示.折断力越大越好.试验的目的就是想寻求提高折断力的新工艺.

在烟灰砖的生产中,成型水分的多少、碾压时间的长短、每次碾压时投多少料都可能对折断力有影响.成型水分、碾压时间、一次碾压料重通常称为因素,以下分别用 A,B,C 记之.对成型水分,要比较三个条件 9%,10%,11%究竟哪一个好?对碾压时间也要比较三个条件 8 分,10 分,12 分究竟哪一个好?对一次碾压的料重,也要比较三个条件 330 kg, 360 kg, 400 kg 究竟哪一个好?每个因素要比较的各个条件称为因素的水平(或叫位极).在这个例子中共有三个因素,每个因素有三个水平.可列表如下:

水平 \ 因素	成型水分	碾压时间	一次碾压料重
1 水平	9%	8 分	330 kg
2 水平	10%	10 分	360 kg
3 水平	11%	12 分	400 kg

通过试验，希望回答下列三个问题：

(1) 对指标（折断力）的影响，哪个因素重要，哪个因素不重要？

(2) 每个因素中哪个水平最好？

(3) 各因素以什么样的水平搭配起来对指标最有利？

这是多因素试验中比较典型的几个问题. 当然，在实际工作中因素不必限于三个，各因素的水平个数不必相等. 怎样安排试验？如果各因素的所有水平组合（在例 1.1 中水平组合共有 27 种）都进行试验，则叫做全面试验. 在实际工作中对于因素个数较少（例如单因素或两因素）的试验常采用全面试验. 当因素较多时，全面试验用得很少，因为工作量太大，试验费用太高. 有时由于试验物资的限制，根本不可能进行全面试验. 一个重要问题是，如何选出一部分有代表性的水平组合来进行试验？这就是试验设计问题. 本章从 §2 起要论述试验设计的方法. 本节只讨论全面试验下的数据分析，着重介绍方差分析法.

首先讲单因素试验.

所谓单因素试验（也叫单因素比较试验），就是在试验中只考虑一个因素，试验的目的在于比较因素各水平上指标值的差别.

例 1.2（饲料对比试验）① 为发展我国机械化养鸡，某研究所根据我国的资源情况，研究用槐树粉、苜蓿粉等原料代替国外用鱼粉做鸡饲料的办法. 他们研究了三种饲料配方：第一种，以鱼粉为主的鸡饲料；第二种以槐树粉、苜蓿粉为主加少量鱼粉；第三种，以槐树粉、苜蓿粉为主加少量化学药品. 后两种是他们研制的新配方. 为比较三种饲料在养鸡增肥上的效果，各喂养 10 只母雏鸡，于 60 天后观测它们的重量. 如下表所示：

---

① 本例取自[7]的第 157 页.

饲　料	鸡　重　(g)									
第一种	1 073	1 058	1 071	1 037	1 066	1 026	1 053	1 049	1 065	1 051
第二种	1 016	1 058	1 038	1 042	1 020	1 045	1 044	1 061	1 034	1 049
第三种	1 084	1 069	1 106	1 078	1 075	1 090	1 079	1 094	1 111	1 092

在这项试验中，指标是 60 天的鸡重，因素是饲料，共取三个水平．试验的目的是要比较三种饲料在养鸡增肥的效果上有何差别．这是单因素试验．

如何进行比较呢？如果不存在试验"误差"，那很简单．只要对每种饲料养一只雏鸡，在 60 天后称一下重量就是了．然而，由于鸡与鸡之间的个体差别造成较大的误差，故所观察到的数据并不能完全反映饲料的结果．自然想到比较各种饲料下的平均鸡重，从表中可算出：第一种饲料下平均鸡重为 1 055 g，第二种饲料下平均鸡重 1 041 g，第三种饲料下的平均鸡重为 1 088 g．从直观上看，三种饲料在增肥效果上有明显差异，而且第三种饲料好．细心的读者会问：怎么知道三个平均数之间的差异主要由于饲料不同产生而不是由于试验误差所造成？换句话说，如何鉴别平均数的差异主要来自"因素"，还是主要来自"误差"？

这需要一种新的统计分析方法——方差分析．

先进行一般性讨论，然后再回到例 1.2 上去．

只考虑一个因素 A，有 s 个水平：$A_1, A_2, \cdots, A_s (s \geqslant 2)$，问：这些水平对指标 y 的作用是否一样？

设 A_i 水平下重复进行了 r 次试验($r \geqslant 2$)，数据是 $y_{i1}, y_{i2}, \cdots, y_{ir}$，我们假定有下列模型：

$$y_{ij} = \mu_i + e_{ij} \quad (i=1,\cdots,s; \quad j=1,\cdots,r) \tag{1.1}$$

其中 $\{e_{ij}\}$ 相互独立，且 $e_{ij} \sim N(0, \sigma^2)$，$\sigma$ 未知．

待检验的假设是：

$$H_A: \mu_1 = \mu_2 = \cdots = \mu_s \tag{1.2}$$

这是线性模型的参数的检验问题．第四章中已有一般理论，现在要用到 H_A 的检验上来．记

$$Y=\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{sr} \end{pmatrix}, \quad X=\begin{pmatrix} E_r & & & \\ & E_r & & \\ & & \ddots & \\ & & & E_r \end{pmatrix}_{rs\times s}$$

（X 的秩是 s），

$$E_r=\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{r\times 1}, \quad e=\begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{sr} \end{pmatrix}$$

易知模型(1.1)可写成

$$Y = X\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_s \end{pmatrix}+e$$

设 $\mu_1,\cdots,\mu_s$ 的最小二乘估计是 $\hat{\mu}_1,\cdots,\hat{\mu}_s$，则残差平方和

$$Q=\left\| Y-X\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_s \end{pmatrix} \right\|^2$$

$$=\min_{\mu_1\cdots\mu_s}\sum_{i,j}(y_{ij}-\mu_i)^2$$

在 H_A 成立下的残差平方和

$$Q_A=\min_{\mu}\sum_{i,j}(y_{ij}-\mu)^2$$

按一般理论，在 H_A 下统计量

$$F=\frac{(Q_A-Q)/(s-1)}{Q/(rs-s)}\sim F(s-1,\ s(r-1))$$

故可用统计量 F 检验假设 H_A.

我们指出有下列简单公式：

$$Q_A-Q=r\sum_i(\bar{y}_{i\cdot}-\bar{y})^2$$

其中

$$\bar{y}_{i\cdot} \stackrel{\mathrm{d}}{=} \frac{1}{r}\sum_{j=1}^{r} y_{ij}, \quad \bar{y} = \frac{1}{rs}\sum_{i,j} y_{ij}$$

实际上，

$$\sum_{i,j}(y_{ij}-\mu_i)^2 = \sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i,j}(\bar{y}_{i\cdot}-\mu_i)^2$$

故

$$Q = \sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2$$

由于

$$\sum_{i,j}(y_{ij}-\mu)^2 = \sum_{i,j}(y_{ij}-\bar{y})^2 + \sum_{i,j}(\bar{y}-\mu)^2$$

易知

$$Q_A = \sum_{i,j}(y_{ij}-\bar{y})^2 = \sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i,j}(\bar{y}_{i\cdot}-\bar{y})^2 \qquad (1.3)$$

于是

$$Q_A - Q = r\sum_{i}(\bar{y}_{i\cdot}-\bar{y})^2$$

所以检验 H_A 的统计量是

$$F = \frac{r\sum_{i}(\bar{y}_{i\cdot}-\bar{y})^2/(s-1)}{\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2/s(r-1)}$$

(1.3)是很重要的平方和分解公式，它把一个平方和分解为两个平方和. 每个平方和都有明确的直观意义：(1.3)式的左端乃是全部原始数据围绕平均值产生的波动平方和，而(1.3)式右端第一项表示由于试验误差引起的波动平方和(同一水平下数据引起的波动平方和)，右端第二项是数据由于因素的不同水平而引起的波动平方和. 很明显，若(1.3)式右端第一项比起第二项大得多，则说明因素的各水平对指标的影响无显著差异，反之，若第二项比第一项大得多，则说明因素的作用很大，各水平的影响有显著差异. 这就给出了上述用统计量 F 进行检验的直观意义.

把全部原始数据产生的波动平方和分解成若干个有特定意义

的平方和(其中有一个是误差引起的平方和),然后将某些平方和与误差引起的平方和进行比较,以对某些"假设"进行检验,这种方法通称方差分析,是 R. A. Fisher 于 1923 年首创的. 在具体应用时,常用方差分析表的格式. 单因素情形下的一元方差分析表如下:

方差来源	平方和	自由度	F 值	显著性
因　素	$S_A=r\sum_i(\overline{y}_{i\cdot}-\overline{y})^2$	$s-1$	$\dfrac{S_A/(s-1)}{S_e/s(r-1)}$	
误　差	$S_e=\sum_{i,j}(y_{ij}-\overline{y}_{i\cdot})^2$	$s(r-1)$		
总　和	$S_T=\sum_{i,j}(y_{ij}-\overline{y})^2$	$rs-1$		

若因素 A 对应的 F 值超过显著性水平 $\alpha=0.05$ 对应的 F 分布临界值,则在"显著性"一栏里记上 *,否则什么也不填. 当 F 值超过 $\alpha=0.01$ 对应的 F 分布临界值时,有时在"显著性"一栏里记上 * *. * 或 * * 表示应该拒绝假设 H_A,此时称因素 A 是显著的. 若没有记号 *,则表示不拒绝假设 H_A,也称因素 A 是不显著的.

现在将上述一般性理论用到例 1.2 上去,得下列方差分析表:

方差来源	平方和	自由度	F 值	显著性
因素(饲料)	11 675	2	28.34	* *
误　　差	5 569	27		
总　　和	17 244	29		

由此可见,因素是显著的,即不同饲料在增肥效果上确有差别. 既然第三种饲料下的平均鸡重高于第一种和第二种饲料下的平均鸡重,故第三种饲料增肥效果最好①.

---

① 当假设(1.2)被拒绝(即因素 A 是显著的)时,只表明 $\mu_1,\cdots,\mu_s$ 不全相等,并未指明某个 μ_i 与某个 μ_j 相等与否. 若要进一步判别 μ_i 与 μ_j 相等与否,则需要学习统计学中的"多重比较的方法"(见[7]的第五章). 本书不介绍这些比较深入的内容.

还需指出的是，在单因素试验中，各水平下的试验次数不必相等. 利用线性模型的理论也可导出相应的统计量与方差分析公式，留给读者作为练习.

现在来研究两个因素的试验.

设因素 A 有 s 个水平：$A_1, \cdots, A_s$；因素 B 有 t 个水平：$B_1, \cdots, B_t$ $(s \geqslant 2,\ t \geqslant 2)$. 对所有的水平组合都进行了 r 次试验 $(r \geqslant 2)$. 设 A 处于 i 水平 B 处于 j 水平下指标的理论值是 μ_{ij}，观测值是 $y_{ij1}, y_{ij2}, \cdots, y_{ijr}$，模型是

$$y_{ijk} = \mu_{ij} + e_{ijk} \tag{1.4}$$

$$(i=1,\cdots,s;\quad j=1,\cdots,t;\quad k=1,\cdots,r)$$

其中 $\{e_{ijk}\}$ 是试验误差，我们假定它们独立同分布，且 $e_{ijk} \sim N(0,\sigma^2)$（$\sigma^2$ 未知）.

为了分析各因素对指标的影响. 令

$$\mu = \frac{1}{st}\sum_{i,j}\mu_{ij}$$

$$\alpha_i = \frac{1}{t}\sum_{j=1}^{t}\mu_{ij} - \mu \quad (i=1,\cdots,s)$$

$$\beta_j = \frac{1}{s}\sum_{i=1}^{r}\mu_{ij} - \mu \quad (j=1,\cdots,t)$$

$$\lambda_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$$

则

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \lambda_{ij} \tag{1.5}$$

$$\sum_{i=1}^{s}\alpha_i = \sum_{j=1}^{t}\beta_j = \sum_{i=1}^{s}\lambda_{ij} = \sum_{j=1}^{t}\lambda_{ij} = 0 \tag{1.6}$$

不难看出，α_i 表示 A 的第 i 个水平的“作用”，β_j 表示 B 的第 j 个水平的“作用”. 通常称 α_i 为水平 A_i 的主效应，β_j 为水平 B_j 的主效应，λ_{ij} 为 A_i 与 B_j 下的交互作用效应. 很明显，若 $\lambda_{ij}=0$（一切 i,j），则

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

这表明在因素 A 的 A_i 水平与因素 B 的 B_j 水平下的理论值(真值)可表示为因素 A 与因素 B 单个"作用"的叠加. 若 $\lambda_{ij}\neq 0$,则从(1.5)知,因素 A 与因素 B 对指标的综合"作用"不能归结为单个因素"作用"的叠加,而因素间有一种内在的联系在起作用(所谓交互作用).

待检验的假设有

$$H_1:\ \alpha_1=\alpha_2=\cdots=\alpha_s=0 \tag{1.7}$$

$$H_2:\ \beta_1=\beta_2=\cdots=\beta_t=0 \tag{1.8}$$

$$H_3:\ \lambda_{11}=\lambda_{12}=\cdots=\lambda_{st}=0 \tag{1.9}$$

H_1 的含义是因素 A 没有影响,H_2 的含义是因素 B 没有影响,H_3 的含义是:A 与 B 的交互作用不存在.

先用线性模型的一般理论导出各个检验统计量的抽象形式,然后利用平方和分解的手段找出统计量的具体表达式.

将 $\{y_{ijk}\}$ 按足标的字典排列法写成一向量

$$Y=\begin{pmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{str} \end{pmatrix}$$

记 $\xi=EY$, ξ 可用参数 $\mu,\alpha_i,\beta_j,\lambda_{ij}$ 表示出来. 记

$$W=\{\xi:\ 参数满足(1.6)\}$$

$$W_1=\{\xi:\ 参数满足(1.6)及(1.7)\}$$

$$W_2=\{\xi:\ 参数满足(1.6)及(1.8)\}$$

$$W_3=\{\xi:\ 参数满足(1.6)及(1.9)\}$$

设 W 的维数是 m, W_i 的维数是 $q_i(i=1,2,3)$,Y 在 W 上之投影为 ξ_0,Y 在 W_i 上之投影为 $\xi_i(i=1,2,3)$.

记 $Q=\|Y-\xi_0\|^2$, $Q_i=\|Y-\xi_i\|^2$, 令

$$F_i=\frac{(Q_i-Q)/(m-q_i)}{Q/(n-m)} \quad (n=str)$$

从第四章知,若 H_i 成立,则 $F_i\sim F(m-q_i,\ n-m)$,故可用统计量

F_i 检验假设 H_i.

剩下的问题是如何计算 m, q_1, q_2, q_3 及 Q, Q_1, Q_2, Q_3.

引理 1.1 设 A 是 $n \times p$ 矩阵，u 是 p 维列向量，若空间 $V=\{\xi: 存在 u, 使得 \xi=Au\}$ 中的每个向量的表示法是唯一的（即对任何 $\xi \in V$，只有一个 u 使得 $\xi=Au$），则 V 的维数是 p.

证明：用反证法易知 A 的各列线性无关，从而 V 的维数是 p.

证毕.

由于 W 中的 ξ 可唯一地由参数 $\mu, \alpha_1, \cdots, \alpha_{s-1}, \beta_1, \cdots \beta_{t-1}, \lambda_{11} \cdots \lambda_{1s-1}, \lambda_{21}, \cdots \lambda_{s-1,t-1}$ 表示出. 故 W 的维数

$$m=1+(s-1)+(t-1)+(s-1)(t-1)=st$$

W_1 中的 ξ 可唯一地由 $\mu, \beta_1, \cdots, \beta_{t-1}, \lambda_{11} \cdots \lambda_{s-1,t-1}$ 表示出，故

$$q_1=1+(t-1)+(s-1)(t-1)=st-(s-1)$$

同理知

$$q_2=1+(s-1)+(s-1)(t-1)=st-(t-1)$$

W_3 中的 ξ 可唯一地由 $\mu, \alpha_1, \cdots, \alpha_{s-1}, \beta_1, \cdots, \beta_{t-1}$，表示出，故

$$q_3=1+(s-1)+(t-1)=s+t-1$$

剩下是求 Q, Q_1, Q_2, Q_3. 用平方和分解法，方法十分初等，但表达式比较繁. 引进下列记号

$$\bar{y}_{ij\cdot}=\frac{1}{r}\sum_{k=1}^{r} y_{ijk}, \quad \bar{y}_{i\cdot\cdot}=\frac{1}{tr}\sum_{j,k} y_{ijk}$$

$$\bar{y}=\bar{y}_{\cdots}=\frac{1}{str}\sum_{i,j,k} y_{ijk}$$

其他记号 $\bar{y}_{\cdot jk}$ $\bar{y}_{\cdot\cdot j}$. 等的含义是类似的，不赘述.

$$\begin{aligned}
S &= \sum_{i,j,k}(y_{ijk}-\mu-\alpha_i-\beta_j-\lambda_{ij})^2 \\
&= \sum_{i,j,k}[(y_{ijk}-\bar{y}_{ij\cdot})^2+(\bar{y}_{ij\cdot}-\mu-\alpha_i-\beta_j-\lambda_{ij})^2] \\
&= \sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\cdot})^2+\sum_{i,j,k}[(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\beta_j-\lambda_{ij})+ \\
&\quad (\bar{y}_{i\cdot\cdot}-\mu-\alpha_i)]^2
\end{aligned}$$

$$=\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\cdot})^2+\sum_{i,j,k}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\beta_j-\lambda_{ij})^2$$

$$+\sum_{i,j,k}(\bar{y}_{i\cdot\cdot}-\mu-\alpha_i)^2$$

$$=\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\cdot})^2+\sum_{i,j,k}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}-\lambda_{ij})^2$$

$$+\sum_{i,j,k}(\bar{y}_{\cdot j\cdot}-\bar{y}-\beta_j)^2+\sum_{i,j,k}(\bar{y}_{i\cdot\cdot}-\bar{y}-\alpha_i)^2+\sum_{i,j,k}(\bar{y}-\mu)^2$$

由此可见

$$Q=\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\cdot})^2$$

$$Q_1=Q+\sum_{i,j,k}(\bar{y}_{i\cdot\cdot}-\bar{y})^2$$

$$Q_2=Q+\sum_{i,j,k}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$$

$$Q_3=Q+\sum_{i,j,k}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2$$

于是

$$S_A\overset{\mathrm{d}}{=\!=}Q_1-Q=tr\sum_{i=1}^{s}(\bar{y}_{i\cdot\cdot}-\bar{y})^2$$

$$S_B\overset{\mathrm{d}}{=\!=}Q_2-Q=sr\sum_{j=1}^{t}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$$

$$S_{A\times B}\overset{\mathrm{d}}{=\!=}Q_3-Q=r\sum_{i,j}(\bar{y}_{ij\cdot}-\bar{y})^2-S_A-S_B$$

最后一式的理由是：

$$\sum_{i,j,k}(\bar{y}_{ij\cdot}-\bar{y})^2=\sum_{i,j,k}[(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})+(\bar{y}_{i\cdot\cdot}-\bar{y})$$

$$+(\bar{y}_{\cdot j\cdot}-\bar{y})]^2$$

$$=\sum_{i,j,k}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2+\sum_{i,j,k}(\bar{y}_{i\cdot\cdot}-\bar{y})^2$$

$$+\sum_{i,j,k}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$$

$$=S_{A\times B}+S_A+S_B$$

现在 $m-q_1=s-1$，$m-q_2=t-1$，$m-q_3=(s-1)(t-1)$，$n-m=st(r-1)$，故有计算公式

$$F_1=\frac{S_A/(s-1)}{Q/st(r-1)}$$

$$F_2=\frac{S_B/(t-1)}{Q/st(r-1)}$$

$$F_3=\frac{S_{A\times B}/(s-1)(t-1)}{Q/st(r-1)}$$

方差分析表如下：

方差来源	平方和	自由度	平均平方和	F 值	显著性
A	S_A	$s-1$	$S_A/(s-1)$	F_1	
B	S_B	$t-1$	$S_B/(t-1)$	F_2	
$A\times B$	$S_{A\times B}$	$(s-1)(t-1)$	$S_{A\times B}/(s-1)(t-1)$	F_3	
误差	Q	$st(r-1)$	$Q/st(r-1)$		
总　　和	S_T	$str-1$			

注意

$$S_A+S_B+S_{A\times B}+Q=\sum_{i,j,k}(y_{ijk}-\bar{y})^2$$

等式右端常记为 S_T.

例 1.3　在某橡胶配方中，考虑了三种不同的促进剂、四种不同分量的氧化锌，同样的配方试验两次，测得了 300% 的定强如下：

定强 \ B:氧化锌 \ A:促进剂	B_1	B_2	B_3	B_4
A_1	31,33	34,36	35,36	39,38
A_2	33,34	36,37	37,39	38,41
A_3	35,37	37,38	39,40	42,44

问：氧化锌，促进剂以及它们的交互作用对定强有无显著影

响？什么是主要矛盾？

这属于两因素全面试验的方差分析范围，方差分析表如下：

方差来源	平方和	自由度	F 值	显著性
A	56.6	2	19.4	* *
B	132.2	3	30.2	* *
$A\times B$	4.7	6	0.55	
误　差	17.5	12		
总　和	211.0	23		

注意 F 分布的临界值是 $F_{0.99}(2,12)=6.9$，$F_{0.99}(3,12)=6.0$，$F_{0.95}(6,12)=3.0$，故 A、B 都是高度显著的(用 * * 表示)，$A\times B$ 是不显著的. 换句话说，促进剂和氧化锌总量都是重要的因素，两者相比氧化锌更重要一些. 这两个因素的交互作用不存在，相应的平方和只是误差的一种反映. 这时，可将该项和误差项合并，相应的自由度也合并，以提高分析的精度. 改进后的方差分析表如下：

方差来源	平方和	自由度	F 值	显著性
A	56.6	2	22.8	* *
B	132.1	3	35.4	* *
误　差	22.2	18		
总　和	211.0	23		

因素 A、B 仍是高度显著的.

以上讨论了单因素和两因素全面试验的方差分析，所用的方法可推广到更多因素的全面试验的数据分析上去，由于公式太繁，使用价值较小，这里不介绍了.

§2　多因素试验设计问题的提法与数学模型

前面说过，在试验工作中若要考察的因素较少，可采用全面试

验，即各因素的所有水平组合都安排试验．当因素较多时，这种全面试验就不合适了，这时只能从各因素的所有水平组合中挑选有代表性的一部分进行试验．怎样挑选呢？每一种挑选就是一种试验设计．

为了研究这个问题，我们需要把试验问题提成恰当的数学模型．

假如我们要考察 m 个因素 F_1（有 s_1 个水平），F_2（有 s_2 个水平），…，F_m（有 s_m 个水平）对某指标 a 的影响．用 $1,2,\cdots,s_j$ 记 F_j 的 s_j 个水平（$s_j\geqslant 2$）．当 F_1 取水平 λ_1，F_2 取水平 λ_2，…，F_m 取水平 λ_m 时指标 a 的真值（理论值）为 $a(\lambda_1,\lambda_2,\cdots,\lambda_m)$，注意真值 $\{a(\lambda_1,\lambda_2,\cdots,\lambda_m):1\leqslant\lambda_j\leqslant s_j,j=1,\cdots,m\}$ 共有 $s_1s_2\cdots s_m$ 个，它们是客观存在的，但我们却观察不到．我们能观察到的是观察值

$$y_{\lambda_1\cdots\lambda_m}=a(\lambda_1,\cdots,\lambda_m)+e_{\lambda_1\cdots\lambda_m}$$

这里 $e_{\lambda_1\cdots\lambda_m}$ 是随机变量，均值是 0，一般假定它服从 $N(0,\sigma^2)$．如何根据观察值 $\{y_{\lambda_1\cdots\lambda_m}\}$ 来估计真值 $a(\lambda_1,\cdots,\lambda_m)$ 或研究其性质，这是很重要的问题．

要对所有的水平组合都做试验取得观察值，那代价太大了，实际工作中常常不允许．故问题的提法是：从所有的 $s_1s_2\cdots s_m$ 个组合中合理地尽量少地挑选一部分组合做试验，然后利用得到的观察值推断出真值 $a(\lambda_1,\cdots,\lambda_m)$ 的值或性质．

真值函数有些什么性质要研究呢？引进必要的定义．令

$$\beta_0=\frac{1}{\prod_{j=1}^{m}s_j}\sum_{\lambda_1,\cdots,\lambda_m}a(\lambda_1,\cdots,a_m)$$

$$\beta_i(\lambda_i)=\frac{s_i}{\prod_{j=1}^{m}s_j}\sum_{\lambda_1,\cdots,\lambda_{i-1},\lambda_{i+1},\cdots,\lambda_m}a(\lambda_1,\cdots,\lambda_m)-\beta_0$$

$$(i=1,\cdots,m,\lambda_i=1,\cdots,s_i)$$

$$\beta_{ij}(\lambda_i,\lambda_j)=\frac{s_i s_j}{\prod_{j=1}^{m} s_j}\sum_{\lambda_1,\cdots,\lambda_{i-1},\lambda_{i+1},\cdots,\lambda_{j-1},\lambda_{j+1},\cdots,\lambda_m} a(\lambda_1,\cdots,\lambda_m)-$$

$$\beta_i(\lambda_i)-\beta_j(\lambda_j)-\beta_0$$

$$(i<j,\quad \lambda_i=1,\cdots,s_4,\quad \lambda_j=1,\cdots,s_j)$$

当 $i>j$ 时,令

$$\beta_{ij}(\lambda_i,\lambda_j)=\beta_{ji}(\lambda_j,\lambda_i)$$

称 β_0 为试验设计问题$\{F_1,\cdots,F_m;s_1,\cdots,s_m\}$的一般平均.

$\beta_i(\lambda_i)$叫做因素 F_i 在水平 λ_i 时的主效应,$\{\beta_i(\lambda_i):\lambda_i=1,\cdots,s_i\}$叫做 F_i 的主效应.

$\beta_{ij}(\lambda_i,\lambda_j)$叫做 F_i 在水平 λ_i,因素 F_j 在水平 λ_j 时交互作用效应. $\{\beta_{ij}(\lambda_i,\lambda_j):\lambda_i=1,\cdots,s_i,\lambda_j=1,\cdots,s_j\}$叫做 F_i 与 F_j 的交互作用效应. 上述效应(包括一般平均)共有 $1+s_1+\cdots+s_m+\sum_{i<j}s_i s_j$ 个,都是试验问题本身固有的,与试验如何设计无关.

注 若对于 $\lambda_i=1,2,\cdots,s_i$ 及 $\lambda_j=1,2,\cdots,s_j$,$s_i\cdot s_j$ 个值$\beta_{ij}(\lambda_i,\lambda_j)$不全相等,则称因素 F_i 和 F_j 间有交互作用,也说 $F_i\times F_j$ 存在. 相反情形(即 $\beta_{ij}(\lambda_i,\lambda_j)\equiv$const(对一切 s_i,s_j)),则称 F_i 与 F_j 间无交互作用,也说 $F_i\times F_j$ 不存在. 二因素间交互作用存在的直观意义是:二因素对指标 a 的综合影响不等于各因素单独影响之和. 例如,在农业试验中考察氮肥(N)和磷肥(P)对大豆产量的影响. 氮肥有两个水平:$N_1=0$(即不施加氮肥),$N_2=6$ 斤(每亩施加氮肥 6 斤①);磷肥有两个水平:$P_1=0$(即不施加磷肥),$P_2=4$(每亩施加 4 斤). 两个因素共有四个水平组合. 在土地情况大体相同的四块试验田上种大豆,每块田对应一种"水平组合",即一种施肥方式. 例如 N_1P_2 对应的施肥方式是,不施氮肥,只施磷肥 4 斤/亩. 设四块田的大豆亩产如下(单位:斤)

---

① 斤为非法定计量单位,1 斤=500 g.

N \ P	$P_1=0$	$P_2=4$
$N_1=0$	400	450
$N_1=6$	430	560

从表中看出，只施 4 斤磷肥，大豆增产 50 斤，这反映磷肥的作用；只施 6 斤氮肥，大豆增产 30 斤，这反映了氮肥的作用；既施 6 斤氮肥又施 4 斤磷肥，大豆增产了 160 斤，它比单独的磷肥的作用与单独的氮肥作用之和要多出 80 斤，这体现了氮肥与磷肥间有交互作用.

类似地，我们还可考虑三个因素间的交互作用，四个因素间的交互作用，…. 大量实际经验表明，在大多数情形下，三个或三个以上因素间的交互作用(称为高级交互作用)都比较小，可以忽略. 甚至某些二因素间的交互作用也可以忽略.

要注意的是，这些效应之间有下列关系式：

$$\sum_{\lambda_i}\beta_i(\lambda_i)=0$$

$$\sum_{\lambda_i}\beta_{ij}(\lambda_i,\lambda_j)=\sum_{\lambda_j}\beta_{ij}(\lambda_i,\lambda_j)=0$$

故独立的参数个数等于

$$1+\sum_{1}^{m}(s_i-1)+\sum_{i<j}(s_i-1)(s_j-1)$$

试验工作者关心下列一个或几个问题：

(1) F_i 是否对指标 a 有影响，即问 F_i 的主效应 $\beta_i(\lambda_i)$ 是否对一切 λ_i 一样大？

(2) F_i 与 F_j 是否有交互作用效应，即问 $\beta_{ij}(\lambda_i,\lambda_j)$ 是否对一切 λ_i,λ_j 一样大？

(3) 怎样估计出 $\beta_i(\lambda_i)$，$\beta_{ij}(\lambda_i,\lambda_j)$？

(4) 如何找出优良的水平组合 $(\lambda_1^*,\lambda_2^*,\cdots,\lambda_m^*)$，即找出 F_1 的水平 λ_1^*，F_2 的水平 λ_2^*，…，F_m 的水平 λ_m^* 使得指标 $a(\lambda_1^*,\lambda_2^*,\cdots,\lambda_m^*)$ 尽可能的大(如果指标 a 的值越大越好的话)或尽可能的小(如果指标 a 的值越小越好的话)？优良的水平组合又称为优良的工程条件(或称生产条件).

上述问题(4)就是工程上的最优化问题，简称全局优化问题．这是实际工作中(特别是工业试验)最常遇到的．而上述问题(1)、(2)、(3)则是科学研究中常见的，旨在探讨各个因素(包括某些交互作用)对指标的影响规律，认识客观事物，也能帮助寻找优良的工程条件．问题(1)、问题(2)和问题(3)合称因素分析问题．全局优化问题与因素分析问题二者有密切联系但侧重点是不同的．何者作为试验工作者的首要目标，对选择试验方案(即试验设计)有很大影响．

大量的实际工作经验表明，若以全局优化问题为首要目标，则在制定试验方案时可以大胆忽略因素间的交互作用，同时安排较多因素进入试验，采用正交设计或均匀设计可使得通过较少的试验次数找出比较优良的工程条件．(当然，所找到的“工程条件”一般不一定是最优的．)若以“因素分析”为首要目标，则在制定试验方案时应考虑因素间的交互作用，特别要注意某些二因素间的交互作用可能存在．此时，正交设计仍是制定试验方案的重要手段．

限于篇幅，我们重点介绍正交设计的某些基本知识．至于均匀设计，可参看方开泰著《均匀设计与均匀设计表》(科学出版社，1994)或[20]．①

§3　正交表和正交设计

正交设计是利用一些特制的正交表来安排试验．我们先介绍

---

①　正交设计和均匀设计都是特殊的试验设计．试验设计的一般定义可如下叙述．设要考虑 m 个因素 $F_1,F_2,\cdots,F_m$ 对某指标的影响，其中 F_j 有 s_j 个水平，分别用 $1,2,\cdots,s_j$ 来表示($j=1,2,\cdots,m$)；所谓一个“试验设计”就是指明 m 个因素的哪些水平组合要做试验，还要说明试验的重复次数．换句话说，对每个水平组合$(\lambda_1,\lambda_2,\cdots,\lambda_m)$规定试验的重复次数 $n(\lambda_1,\lambda_2,\cdots,\lambda_m)$(非负整数)．当 $n(\lambda_1,\lambda_2,\cdots,\lambda_m)=0$ 时意味着相应的水平组合$(\lambda_1,\lambda_2,\cdots,\lambda_m)$不进行试验．从数学上看，一个试验设计(或试验方案)就是一个非负整值映射 $n(\cdot,\cdots,\cdot)$．若 $n(\lambda_1,\cdots,\lambda_m)\leqslant 1$，则称试验是无重复的．

正交表.

设 $m \geqslant 2, s_1, s_2, \cdots, s_m$ 是 m 个不小于 2 的正整数.

定义 3.1 称矩阵 $A=(\lambda_{ij})_{n\times m}$ 是 $s_1\times s_2\times\cdots\times s_m$ 型正交表(记为 $L_n(s_1\times\cdots\times s_m)$),若它满足下列条件:

(1) $\lambda_{ij}\in\{1,2,\cdots,s_j\}$ $(1\leqslant i\leqslant n,\quad 1\leqslant j\leqslant m)$

(2) 对任何 $j_1<j_2, u\in\{1,2,\cdots,s_{j_1}\}, v\in\{1,2,\cdots,s_{j_2}\}$ 有

$$\#\{i:(\lambda_{ij_1},\ \lambda_{ij_2})=(u,v)\}=\frac{n}{s_{j_1}\cdot s_{j_2}}$$

这里 $\# A$ 表示集合 A 的元素个数. 当 $s_1=s_2=\cdots=s_m=s$ 时,正交表记为 $L_n(s^m)$,当 $s_2=s_3=\cdots=s_m=s$ 时正交表记为 $L_n(s_1\times s^{m-1})$.

常用的正交表有 $L_4(2^3)$, $L_8(2^7)$, $L_9(3^4)$, $L_{16}(2^{15})$, $L_{16}(4^5)$, $L_{18}(2\times 3^7)$, $L_{27}(3^{13})$等,下面介绍 $L_4(2^3)$, $L_8(2^7)$ 和 $L_9(3^4)$,其余的表可参看[1]和[5].

表 $L_4(2^3)$是

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$$

表 $L_8(2^7)$是

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 2 & 1 & 2 & 1 \\ 2 & 2 & 1 & 1 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 & 1 & 2 \end{pmatrix}$$

表 $L_9(3^4)$是这样的:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 \\ 2 & 1 & 2 & 3 \\ 2 & 2 & 3 & 1 \\ 2 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \\ 3 & 2 & 1 & 3 \\ 3 & 3 & 2 & 1 \end{pmatrix}$$

构造正交表的方法多种多样，有些很容易，例如 $L_8(2^7)$，$L_{16}(2^{15})$，$L_9(3^4)$；有些则比较复杂，例如 $L_{12}(2^{11})$. 有些参数对应的正交表是不存在的，例如没有 $L_{36}(6^4)$. 构造正交表的一般理论见[7].

我们指出，正交表经过下列三种初等变换仍然是正交表：① 任意两行互换，② 任意两列互换，③ 任何一列中将水平代号进行置换. 因此同型的正交表是很多的. 例如 $L_8(2^7)$有很多张，上面所列的是其中一张(叫做标准表). 值得注意的是，可以证明所有的 $L_8(2^7)$经过上述三种变换可变成上面介绍的那一张，但 $L_{16}(2^{15})$就没有这种唯一性.

如何利用正交表安排试验(即进行正交设计)呢？我们只介绍不考虑因素间交互作用的情形，此时方法十分简单. 至于考虑交互作用的情形如何进行正交设计，方法比较复杂，我们将在本节的末尾略作介绍，以引起读者的注意. 若要更多了解考虑交互作用时的正交设计，请参看[20]、[21]或[7].

从一个具体例子谈起.

例 3.1 烟灰砖试验的正交设计.

继续例 1.1 的讨论. 对因素和水平均用符号来表示. 共有三个因素：A(成型水分)、B(碾压时间)、C(一次碾压料重)，每个因素取三个水平. A 的三个水平是：A_1(成型水分 9%)，A_2(成型水

分 10%），A_3（成型水分 11%）；B 的三个水平是：B_1（碾压时间 8′），B_2（碾压时间 10′），B_3（碾压时间 12′）；C 的三个水平是：C_1（一次碾压 330 kg），C_2（一次碾压 360 kg），C_3（一次碾压 400 kg）.

所有的水平组合有 27 种，如何选出有代表性的一小部分（比如 9 种）进行试验？可用正交表 $L_9(3^4)$ 来实现. 将三个因素 A、B、C 随便放在 $L_9(3^4)$ 的三个列上，例如依次放在第 1、2、3 列上，则就得到一张试验计划表：只要记住因素所在的列中的数字代表该因素的水平. 例如第 1 号试验从表的第 1 行 1 1 1 得到，即因素 A 取 1 水平即成型水分 9%，因素 B 取 1 水平即碾压时间取 8 分钟，因素 C 取 1 水平即一次碾压的料重 330 kg；余类推. 一共 9 次试验，这就是从 27 种水平组合中选出的有代表性部分.

试验计划与试验结果见表 3.1.

表 3.1

列号 / 试验号	A （成型水分） 1	B （碾压时间） 2	C （一次碾压料重） 3	折断力
1	1(9%)	1(8′)	1(330 kg)	16.9
2	1(9%)	2(10′)	2(360 kg)	19.1
3	1(9%)	3(12′)	3(400 kg)	16.7
4	2(10%)	1(8′)	2(360 kg)	19.8
5	2(10%)	2(10′)	3(400 kg)	23.7
6	2(10%)	3(12′)	1(330 kg)	19.0
7	3(11%)	1(8′)	3(400 kg)	25.0
8	3(11%)	2(10′)	1(330 kg)	20.4
9	3(11%)	3(12′)	2(360 kg)	23.1

怎样分析试验结果呢？首先想到直观分析法. 以成型水分为

例，可将三个水平下的平均折断力分别算出. 1 水平下的平均折断力是 17.6，2 水平下是 20.8，3 水平下是 22.8，这就是说，成型水分用 11%最好，最好的与最差的相差 5.3(极差)；对于碾压时间，三个水平下的平均折断力是 20.7、21.1、19.6，故碾压时间取 10 分最好，极差是 1.5；对于一次碾压的料重，三个水平下的平均折断力是 18.8、20.7、21.9，可见一次投料 400 kg 最好，极差是3.1.

综上所述，有下列三条结果.

(1) 成型水分影响最大，其次是一次碾压的料重，再次是碾压时间.

(2) 成型水分是 11%好，碾压时间是 10 分钟好，一次料重是 400 kg 好.

(3) 最好的工艺条件是：成型水分取 11%，碾压时间取 10 分钟，一次料重取 400 kg.

值得注意的是，所指出的最好工艺条件并未在所进行的 9 次试验中出现.

读者自然提出问题：$L_9(3^4)$表的好处在哪里？上述分析计算的理论根据何在？

下面对多因素试验进行一般性讨论. 首先介绍如何用正交表安排试验，然后讨论数据的分析方法与理论基础. 本讲义只讨论最简单、也最有实用价值的情形——可加模型的正交设计.

什么是可加模型呢？考虑因素 F_1(有 s_1 个水平)，F_2(有 s_2 个水平)，…，F_m(有 s_m 个水平)对指标 a 的影响，用 $a(\lambda_1,\cdots,\lambda_m)$表示 F_1 取水平 λ_1，F_2 取水平 λ_2，…，F_m 取水平 λ_m 时的真值，若 $a(\lambda_1,\lambda_2,\cdots,\lambda_m)$有下列分解式：

$$a(\lambda_1,\lambda_2,\cdots,\lambda_m)=a_0+\sum_{j=1}^{m}a_j(\lambda_j) \tag{3.1}$$

则称因素 $F_1,F_2,\cdots,F_m$ 适合可加模型. 可加模型的直观意义是：各因素对指标 a 的影响是各个因素分别影响的叠加，因素间的交

互影响可以忽略. 可加模型当然是一种特殊的模型[①],但实践经验表明,它广泛地存在于实际工作中,而且即使实际情况不完全适合可加模型,但采用可加模型的正交设计安排试验常常也可得出有重要价值的试验结果,特别是能找到比较优良的工程条件.

要注意的是,若分解式(3.1)存在,则它不唯一,但在下列条件下就唯一了.

$$\sum_{\lambda_j=1}^{s_j} a_j(\lambda_j)=0 \quad (j=1,2,\cdots,m) \tag{3.2}$$

满足(3.2)的分解式(3.1)叫做标准分解式. 此时 a_0 正好是一般平均 β_0, $a_j(\lambda_j)$正好是因素 F_j 在水平 λ_j 时的主效应 $\beta_j(\lambda_j)$. 以下只考虑标准分解式.

可加模型下的正交设计是这样的. 设因素 $F_1,F_2,\cdots,F_m$ 分别有 $s_1,s_2,\cdots,s_m$ 个水平. F_j 的各水平由 $1,\cdots,s_j$ 记之. 选一正交表,将 $F_1,F_2,\cdots,F_m$ 任意放于它的 m 个列上,但规定 F_j 必须放在水平数是 s_j 的列上,$j=1,2,\cdots,m$. 将这些列保留,其他的列去掉,得到 $L_n(s_1\times\cdots\times s_m)$型正交表 $\Lambda=(\lambda_{ij})$,于是 Λ 的各行就代表各号试验条件. 例如 Λ 的第 i 号试验条件(水平组合)是:因素 F_1 取 λ_{i1}水平,F_2 取 λ_{i2}水平,$\cdots$,F_m 取 λ_{im}水平. 换句话说,Λ 实现了试验设计,Λ 叫做设计阵. 从上述可见,对可加模型来讲,只要找到了合适的正交表,进行正交设计是十分容易的. 把各因素往表的各列上放,有相当大的任意性.

现在来研究,在可加模型的假定下,如何根据正交设计得到的试验数据进行分析. 设各号试验条件下均进行了 r 次试验

---

① 比可加模型更复杂的是带简单交互作用的模型:

$$a(\lambda_1,\cdots,\lambda_m)=a_0+\sum_{j=1}^{m}a_j(\lambda_j)+\sum_{i\neq j}a_{ij}(\lambda_i,\lambda_j)$$

本讲义不讨论这种模型下的正交设计.所谓"不考虑交互作用时的正交设计"就是指当作可加模型时进行的正交设计.

$(r\geqslant 1)$，第 i 号试验条件下的数据是 $y_{ik}(i=1,\cdots,n,k=1,\cdots,r)$. 于是有数学模型：

$$y_{ik}=\beta_0+\sum_{j=1}^{m}\beta_j(\lambda_{ij})+e_{ik}\quad(i=1,\cdots,n,\quad k=1,\cdots,r)\qquad(3.3)$$

其中 $\Lambda=(\lambda_{ij})$ 是设计阵，e_{ik} 是试验误差，我们假定 $\{e_{ik}\}$ 是相互独立的，期望是 0，方差相等. 有时进一步假定 $e_{ik}\sim N(0,\sigma^2)$. 注意(3.3)中的未知参数满足约束条件：

$$\sum_{\lambda_j=1}^{s_j}\beta_j(\lambda_j)=0\quad(j=1,2,\cdots,m)\qquad(3.4)$$

我们的第一项任务是根据试验数据估计未知参数 $\{\beta_0,\beta_j(\lambda_j)\}$ 的值，从而可估计出 $a(\lambda_1,\cdots,\lambda_m)$ 的值.

(3.3)乃是带有约束条件的线性模型，我们可以利用第四章中的一般理论. 令

$$y_i=\bar{y}_{i\cdot}=\frac{1}{r}\sum_{k=1}^{r}y_{ik}\,,\ \varepsilon_i=\frac{1}{r}\sum_{k=1}^{r}e_{ik}\quad(i=1,2,\cdots,n)$$

$$Y=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\quad \beta_j=\begin{pmatrix}\beta_j(1)\\ \vdots\\ \beta_j(s_j)\end{pmatrix},\quad j=1,2,\cdots,m$$

$$\varepsilon=\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix},\quad E_s=\begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix}\quad(s\text{ 个 }1)$$

$$\beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_m\end{pmatrix},\quad x_{ik}^{(j)}=\delta(\lambda_{ij},k)$$

$$X_j=(x_{ik}^{(j)'})_{n\times s_j}$$

这里

$$\delta(u,v)=\begin{cases}1, & u=v\\ 0, & u\neq v\end{cases}$$

$$X=(E_n,X_1,\cdots,X_m)$$

于是有

$$Y=X\cdot\beta+\varepsilon$$

这里 X 被设计阵 Λ 唯一确定,叫做 Λ 的从属矩阵,记作 $X(\Lambda)$. 由于 β 受到约束,我们设法引进新的参数,它们是无约束的.

设 Γ_j 是 s_j 阶正交矩阵且有结构:

$$\Gamma_j=\left(\frac{1}{\sqrt{s_j}}E_{s_j},T_j\right),\quad j=1,\cdots,m$$

令 $\eta_j=T_j'\beta_j$, $\eta_0=\beta_0$. 于是 $T_j\eta_j=T_jT_j'\beta_j=\left(I_{s_j}-\frac{1}{s_j}E_{s_j}\cdot E_{s_j}'\right)\beta_j=\beta_j$, 这里利用了关系式 $E_{s_j}'\beta_j=0$(见(3.4)). 于是

$$X\beta=(E_n,X_1,\cdots,X_m)\begin{pmatrix}\eta_0\\ T_1\eta_1\\ \vdots\\ T_m\eta_m\end{pmatrix}$$

$$=(E_n,X_1T_1,\cdots,X_mT_m)\begin{pmatrix}\eta_0\\ \eta_1\\ \vdots\\ \eta_m\end{pmatrix}$$

以下记

$$\widetilde{X}=(E_n, X_1T_1, \cdots, X_mT_m)$$

$$\eta=\begin{pmatrix}\eta_0\\ \eta_1\\ \vdots\\ \eta_m\end{pmatrix}$$

于是有

$$Y=\widetilde{X}\eta+\varepsilon$$

这个模型里参数 η 是自由的.

* **引理 3.1** $R(\widetilde{X})=R(X)$(这里 $R(A)$是 A 的秩).

证明：令

$$\Gamma=\begin{pmatrix}1 & & & 0\\ & \Gamma_1 & & \\ & & \ddots & \\ 0 & & & \Gamma_m\end{pmatrix}$$

则 Γ 是满秩方阵. 于是 $R(X)=R(X\Gamma)$. 由于 $X\Gamma=(E_n, X_1\Gamma_1, \cdots, X_m\Gamma_m)$，而 $X_j\Gamma_j=\left(\frac{1}{\sqrt{s_j}}E_n, X_jT_j\right)$. 可见 $X\Gamma$ 与$(E_n, X_1T_1, \cdots, X_mT_m)=\widetilde{X}$ 有相同的秩. 证毕.

* **引理 3.2** 设 $\Lambda=(\lambda_{ij})$是 $L_n(s_1\times\cdots\times s_n)$型正交表，则

$$X'_iX_j=\frac{n}{s_is_j}E_{s_i}E'_{s_j} \quad (i\neq j)$$

$$X'_jX_j=\frac{n}{s_j}I_{s_j} \quad (I_s \text{ 是 } s \text{ 阶单位阵})$$

证明：从定义直接推知.

* **引理 3.3** 设 Λ 是正交设计，则 $\widetilde{X}$ 的秩等于

$$1+\sum_1^m(s_j-1)$$

证明：因为 $X_jE_{s_j}=E_n$，从引理 3.2 知

$$E'_nX_jT_j=\frac{n}{s_j}E'_{s_j}T_j=0$$

又 $i \neq j$ 时

$$T'_i X'_i X_j T_j = \frac{n}{s_i s_j} T'_i E_{s_i} E'_{s_j} T_j = 0$$

另一方面

$$T'_j X'_j X_j T_j = \frac{n}{s_j} T'_j T_j = \frac{n}{s_j} I_{s_j - 1}$$

于是有

$$\widetilde{X}'\widetilde{X} = \begin{pmatrix} n & & & & 0 \\ & \frac{n}{s_1} I_{s_1-1} & & & \\ & & \frac{n}{s_2} I_{s_2-1} & & \\ & & & \ddots & \\ 0 & & & & \frac{n}{s_m} I_{s_m-1} \end{pmatrix}$$

可见

$$R(\widetilde{X}) = R(\widetilde{X}'\widetilde{X}) = 1 + \sum_1^m (s_j - 1) \qquad \text{证毕.}$$

* **系 3.1** 设 Λ 是 $L_n(s_1 \times \cdots \times s_m)$ 型正交表，则有重要不等式：

$$\sum_1^m (s_j - 1) \leqslant n - 1$$

证明：从引理 3.3 直接推出.

η 的最小二乘估计是

$$\hat{\eta} = (\widetilde{X}'\widetilde{X})^{-1} \widetilde{X}'Y$$

于是 β 的最小二乘估计是

$$\hat{\beta} = \begin{pmatrix} 1 & & & & \\ & T_1 & & & \\ & & T_2 & & \\ & & & \ddots & \\ & & & & T_m \end{pmatrix} \cdot \hat{\eta}$$

于是 $\hat{\beta}_j = T_j \hat{\eta}_j \ (j \geqslant 1)$，$\hat{\beta}_0 = \hat{\eta}_0$，不难看出

$$(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}' = \begin{pmatrix} \frac{1}{n}E'_n \\ \frac{s_1}{n}T'_1X'_1 \\ \vdots \\ \frac{s_m}{n}T'_mX'_m \end{pmatrix}$$

$$\hat{\beta}_0 = \frac{1}{n}E'_nY = \bar{y} = \frac{1}{nr}\sum_{i,k} y_{ik}$$

$$\begin{aligned} \hat{\beta}_j &= \frac{s_j}{n}T_jT'_jX'_jY \\ &= \frac{s_j}{n}\left(I_{s_j} - \frac{1}{s_j}E_{s_j}E'_{s_j}\right)X'_jY \\ &= \frac{s_j}{n}X'_jY - \frac{1}{n}E_{s_j}(X_jE_{s_j})'Y \\ &= \frac{s_j}{n}X'_jY - E_{s_j}\bar{y} \end{aligned}$$

记 $$B_{j\lambda} = \sum_{i:\lambda_{ij}=\lambda}\sum_{k=1}^{r} y_{ik} \quad (\lambda = 1, 2, \cdots, s_j)$$

则 $$B_{j\lambda} = \sum_{i=1}^{n}\delta(\lambda_{ij}, \lambda)\sum_{k=1}^{r} y_{ik}$$

于是

$$X'_jY = \frac{1}{r}\begin{pmatrix} B_{j1} \\ \vdots \\ B_{js_j} \end{pmatrix}$$

故

$$\hat{\beta}_j = \frac{s_j}{nr}\begin{pmatrix} B_{j1} \\ \vdots \\ B_{js_j} \end{pmatrix} - \bar{y}E_{s_j}$$

$$\hat{\beta}_j(\lambda) = \frac{s_j}{nr}B_{j\lambda} - \bar{y} \quad (\lambda = 1, \cdots, s_j) \tag{3.5}$$

注意 $B_{j\lambda}$ 乃是 F_j 的 λ 水平下全部数据之和，$\frac{nr}{s_j}$ 乃是该水平的出现次数，因而 $\frac{s_j}{nr}B_{j\lambda}$ 乃是 λ 水平下的平均值．故公式(3.5)的意义在直观上是相当明显的．

真值 $a(\lambda_1,\cdots,\lambda_m)$ 的估计值是

$$\hat{a}(\lambda_1,\cdots,\lambda_m)=\hat{\beta}_0+\sum_1^m\hat{\beta}_j(\lambda_j)$$

$$=\sum_{j=1}^m\frac{s_j}{nr}B_{j\lambda_j}-(m-1)\bar{y}$$

如果指标值越大越好，取 $\lambda_j^{(0)}$ 使得 $\hat{\beta}_j(\lambda_j^{(0)})=\max\limits_{\lambda}\hat{\beta}_j(\lambda)$，则 $\lambda_j^{(0)}$ 就是因素 F_j 的最好水平．显然

$$\frac{s_j}{nr}B_{j\lambda_j^{(0)}}=\max_{\lambda}\frac{s_j}{nr}B_{j\lambda}$$

可见，找最好的水平等价于找平均值最大的水平．当然 $(\lambda_1^{(0)},\cdots,\lambda_m^{(0)})$ 是最好的工程条件．

在上面的分析中有一个重要的缺点：未分析试验误差的大小．试验数据总含有误差，对因素 F_j 而言，各水平下效应估计值 $(\hat{\beta}_{j\lambda})$ 的差异主要由于水平不同所产生，还是主要由于试验误差所产生？换句话说，应该检验下列假设：

$$H_j:\beta_j(1)=\cdots=\beta_j(s_j)=0 \tag{3.6}$$

当 H_j 被拒绝时（这时称 F_j 是显著的），才能说 F_j 的各水平对指标的影响有差异，哪个水平最好．

怎样检验 H_j 呢？这是比较复杂的问题．通常有两种处理方法．

（一）直观分析法

试验工作者根据已往的经验对本项试验中的误差有大致的估计，可以判断 F_j 的各水平效应估计值 $\hat{\beta}_j(1),\cdots,\hat{\beta}_j(\lambda_j)$ 的差异是否主要来自试验误差，根据经验来拒绝还是接受 H_j．当 H_j 被拒绝

时，挑选 F_j 的最好的水平；当 H_j 被接受时，原则上可随便挑选 F_j 的水平（当然这时应根据实际工作的其他需要来挑选）. 这种凭经验挑选出的水平组合（例如（$\lambda_1^{(0)},\cdots,\lambda_m^{(0)}$））究竟是不是最好的“工程条件”？通常要进行验证试验：对所选出水平组合（例如（$\lambda_1^{(0)},\cdots,\lambda_m^{(0)}$））补充进行一次或多次重复试验，看看这时的指标值是否达到了优良的水平（注意，不是验证“最好”，而是验证是否达到了实际工作所要求的“优良”）.

（二）方差分析法

这是一种比较精确的数学方法，我们在全面试验的数据分析中已介绍过. 使用此法的困难在于：如何对试验误差给出客观的较精确的估计. 我们指出，要得到试验误差的客观估计，最好是进行重复试验，即各号试验条件下至少要做两次试验.

怎样进行方差分析呢？设在第 i 号试验条件 $\lambda_{i1},\cdots,\lambda_{im}$ 下的试验数据为 $y_{i1},\cdots,y_{ir}$. 在数学模型（3.3）中，我们假定 $\{e_{ik}\}$ 独立同分布且 $e_{ik}\sim N(0,\sigma^2)$（σ^2 未知）.

沿用前面的记号 $B_{j\lambda},\hat{\beta}_0,\hat{\beta}_j(\lambda_j)$ 及 $\hat{\eta}_j(\lambda_j)$ 等. 令

$$S_{F_j}=\frac{s_j}{nr}\sum_{\lambda=1}^{s_j}B_{j\lambda}^2-\frac{T^2}{nr}\quad\left(T=\sum_{i,k}y_{ik}\right)\tag{3.7}$$

$$y_i=\frac{1}{r}\sum_{k=1}^{r}y_{ik}$$

$$\hat{\varepsilon}_i=y_i-\hat{\beta}_0-\sum_{j=1}^{m}\hat{\beta}_j(\lambda_{ij})\quad（残差）,i=1,\cdots,n$$

$$S_\varepsilon=\hat{\varepsilon}'\hat{\varepsilon}\quad(\hat{\varepsilon}=(\hat{\varepsilon}_1,\cdots,\hat{\varepsilon}_n)')$$

$$\bar{y}=\frac{T}{nr}\quad（全部数据的平均值）$$

我们首先证明下列恒等式

$$\sum_1^n(y_i-\bar{y})^2=\frac{1}{r}\sum_{j=1}^{m}S_{F_j}+S_\varepsilon\tag{3.8}$$

实际上，$Y=\widetilde{X}\hat{\eta}+\hat{\varepsilon},\widetilde{X}\hat{\eta}\perp\hat{\varepsilon}$，于是有

$$Y'Y=\hat{\eta}'\widetilde{X}'\widetilde{X}\hat{\eta}+\hat{\varepsilon}'\hat{\varepsilon}$$

$$=\hat{\eta}'\begin{pmatrix} n & & & \\ & \frac{n}{s_1}I_{s_1-1} & & \\ & & \ddots & \\ & & & \frac{n}{s_m}I_{s_m-1} \end{pmatrix}\hat{\eta}+S_\varepsilon$$

$$=n(\hat{\eta}_0)^2+\sum_{j=1}^{m}\frac{n}{s_j}\hat{\eta}'_j\hat{\eta}_j+S_\varepsilon$$

但 $\hat{\eta}_0=\bar{y}$,故

$$\sum_{1}^{n}(y_i-\bar{y})^2=\sum_{1}^{m}\frac{n}{s_j}\hat{\eta}'_j\hat{\eta}_j+S_\varepsilon$$

我们指出

$$\frac{n}{s_j}\hat{\eta}'_j\hat{\eta}_j=\frac{1}{r}S_{F_j} \tag{3.9}$$

实际上,

$$\hat{\eta}'_j\hat{\eta}_j=\left(\frac{s_j}{n}T'_jX'_jY\right)'\left(\frac{s_j}{n}T'_jX'_jY\right)$$

$$=\frac{s_j^2}{n^2}Y'X_jT_jT'_jX'_jY$$

$$=\frac{s_j^2}{n^2}Y'X_j\left(I_{s_j}-\frac{1}{s_j}E_{s_j}E'_{s_j}\right)X'_jY$$

$$=\frac{s_j^2}{n^2}\left[Y'X_jX'_jY-\frac{1}{s_j}(E'_{s_j}X'_jY)^2\right]$$

$$=\frac{s_j^2}{n^2r^2}\left[\begin{pmatrix} B_{j1} \\ \vdots \\ B_{js_j} \end{pmatrix}'\begin{pmatrix} B_{j1} \\ \vdots \\ B_{js_j} \end{pmatrix}-\frac{1}{s_j}\left(\sum_{\lambda=1}^{s_j}B_{j\lambda}\right)^2\right]$$

$$=\frac{s_j^2}{n^2r^2}\left[\sum_{\lambda=1}^{s_j}B_{j\lambda}^2-\frac{T^2}{s_j}\right]$$

故(3.9)成立,从而(3.8)成立.

由于诸 $\hat{\eta}_i$ 相互独立且与 S_ε 独立，所以 $S_{F_1},\cdots,S_{F_m},S_\varepsilon$ 相互独立.

利用关系式

$$\sum_{i,k}(y_{ik}-\bar{y})^2=\sum_{i,k}(y_{ik}-y_i)^2+r\sum_i(y_i-\bar{y})^2$$

从(3.8)得下列方差分解式

$$\sum_{i,k}(y_{ik}-\bar{y})^2=Q+\sum_{j=1}^{m}S_{F_j}+rS_\varepsilon \tag{3.10}$$

其中

$$Q=\sum_{i,k}(y_{ik}-y_i)^2 \quad \left(y_i=\frac{1}{r}\sum_{k=1}^{r}y_{ik}\right)$$

是误差平方和.

由于 y_i 与 $\sum_{k=1}^{r}(y_{jk}-y_j)^2$（一切 i,j）相互独立，不难推知 Q，$S_{F_1},\cdots,S_{F_m},S_\varepsilon$ 相互独立. 显然

$$\frac{Q}{\sigma^2}=\sum_{i=1}^{n}\frac{1}{\sigma^2}\sum_{k=1}^{r}(y_{ik}-y_i)^2\sim\chi^2(n(r-1))（当 r\geqslant 2）$$

由于

$$\hat{\eta}_j\sim N\left(\eta_j,\frac{s_j}{n}\cdot\frac{\sigma^2}{r}I_{s_j-1}\right)$$

故

$$\frac{nr}{s_j\sigma^2}(\hat{\eta}_j-\eta_j)'(\hat{\eta}_j-\eta_j)\sim\chi^2(s_j-1).$$

在假设 H_j（见(3.6)）下

$$\frac{nr}{s_j\sigma^2}\hat{\eta}_j'\hat{\eta}_j\sim\chi^2(s_j-1)$$

从而

$$\frac{1}{\sigma^2}S_{F_j}\sim\chi^2(s_j-1)$$

令

$$F^{(j)}=\frac{S_{F_j}/(s_j-1)}{Q/n(r-1)}$$

在 H_j 下 $F^{(j)} \sim F(s_j-1, n(r-1))$，故可用统计量 $F^{(j)}$ 来检验假设 H_j.

在上面介绍的方差分析法中要求各号试验条件下均进行过重复试验（重复数 $r \geqslant 2$）. 如果不作重复试验，每号试验条件下只有一个试验数据（分别用 $y_1, \cdots, y_n$ 表示），怎样对假设(3.6)进行检验？这是实际工作者常遇到的问题. 这时对试验误差难以作出确切的估计，通常可用下列无重复试验的方差分析法进行分析.

在(3.8)中，令 $r=1$ 得

$$S_\varepsilon = \sum_1^n (y_i - \bar{y})^2 - \sum_{j=1}^m S_{F_j} \tag{3.11}$$

这里 S_ε 是残差平方和. 根据线性模型的一般理论，当 $f = n - \left(1 + \sum_1^m (s_j - 1)\right) > 0$ 时

$$\frac{S_\varepsilon}{\sigma^2} \sim \chi^2(f)$$

于是在 H_j（见(3.6)）下

$$F = \frac{S_{F_j}/(s_j - 1)}{S_\varepsilon / f} \tag{3.12}$$

服从 $F(s_j - 1, f)$.

故可用统计量(3.12)检验假设 H_j.

如果检验的结果是拒绝 H_j，则表明因素 F_j 是显著的（即它的不同水平中至少有两个对指标的影响有差异）. 如果检验的结果不能拒绝 H_j，而自由度 f 又相当大，则表明因素 F_j 的不同水平对指标的影响无差异. 当自由度 f 比较小又不能拒绝 H_j 时，只表明：从试验数据来看未发现 F_j 的不同水平对指标有不同影响.

$L_n(s_1 \times \cdots \times s_m)$ 型正交设计情形下（无重复试验）的方差分析表如下：

方差来源	平方和	自由度	F 值	显著性
因素 F_1	S_{F_1}	s_1-1	$\dfrac{S_{F_1}/(s_1-1)}{S_s/f}$	
F_2	S_{F_2}	s_2-1	$\dfrac{S_{F_2}/(s_2-1)}{S_s/f}$	
	⋮	⋮	⋮	
F_m	S_{F_m}	s_m-1	$\dfrac{S_{F_m}/(s_m-1)}{S_s/f}$	
误　差	S_ε	$f=n-1-\sum\limits_1^m(s_j-1)$		
总　和	S_T			

表中因素 F_j 的水平数是 $s_j(i=1,\cdots,m)$，S_{F_j} 的计算公式见(3.7)，S_s 的计算公式见(3.11).

$$S_T \stackrel{\mathrm{d}}{=} \sum_1^n (y_i-\bar{y})^2 = \sum_1^n y_i^2 - \frac{\left(\sum\limits_1^n y_i\right)^2}{n}$$

例 3.2　烟灰砖试验的方差分析.

继续例 3.1 的讨论. 根据已往的实践经验，可以认为所考虑的三个因素对折断力的影响适合可加模型. 根据前面给出的试验方案及试验数据，得方差分析表如下：

方差来源	平方和	自由度	F 值	显著性
A	43.89	2	4.46	
B	3.46	2	0.35	
C	14.96	2	1.52	
误　差	9.85	2		
总　和	72.13	8		

查 F 分布表知临界值 $F_{0.95}(2,2)=19.0$，故在显著性水平 $\alpha=0.05$ 时三个因素都不是显著的. 由于误差的自由度太小，这只表

明从本试验的 9 个数据来看未发现因素的不同水平对折断力有不同影响.

但是,根据以往类似的试验工作的经验,同一条件下试验两次,折断力指标一般相差不超过 1. 采用直观分析法,直接对三个因素的不同水平下的指标平均值进行比较,可以认为三个因素都对折断力有影响,最好的工艺条件就是以前说过的:成型水分 11%,碾压时间 10 分,一次料重 400 kg.

例 3.3 某化工厂生产一种试剂,长期以来收率一直不高,只有 65%左右. 为了提高收率,想进行一些试验,认为下列几个因素可能对收率影响较大:反应温度、反应时间、硫酸浓度、硫酸产地、操作方式,每个因素想比较两个不同的条件(水平),具体是:

A(反应温度):A_1(50 ℃),A_2(70 ℃)

B(反应时间):B_1(1 小时),B_2(2 小时)

C(硫酸浓度):C_1(17%),C_2(27%)

D(硫酸产地):D_1(天津),D_2(上海)

E(操作方式):E_1(搅拌),E_2(不搅拌)

这是五因素、每因素二水平的试验,可用正交表 $L_8(2^7)$来安排. 例如将 A,B,C,D,E 依次放在该表的第 1,2,3,4,7 各列的上方,即得试验计划. 现将试验计划和按计划进行试验得到的数据以及分析计算过程写在同一张表(表 3.2)上.

表 3.2 的意义如下:试验号 1 至 8 表示 8 次试验的具体内容,例如第 4 号试验对应的行是 1、2、2、2、1,即因素 A 取 A_1 水平,B 取 B_2 水平,C 取 C_2 水平,D 取 D_2 水平,E 取 E_1 水平. Ⅰ表示各列中 1 对应的各数据之和,Ⅱ表示各列中 2 对应的各数据之和,各因素所在列的 $\frac{\text{I}}{4}-\hat{\mu}$,$\frac{\text{II}}{4}-\hat{\mu}$ 分别是该因素 1 水平与 2 水平的效应. 各因素所在列的下方的 S 是该因素的水平变动产生的数据的波动平方和. 以因素 A 为例,从(3.7)知

表 3.2

行号 \ 列号	A 1	B 2	C 3	D 4	5	6	E 7	y 数据(%)
1	1	1	1	1	1	1	1	65
2	1	1	1	2	2	2	2	74
3	1	2	2	1	1	2	2	71
4	1	2	2	2	2	1	1	73
5	2	1	2	1	2	1	2	70
6	2	1	2	2	1	2	1	73
7	2	2	1	1	2	2	1	62
8	2	2	1	2	1	1	2	67
Ⅰ	283	282	268	268			273	$\hat{\mu}=\bar{y}=\frac{555}{8}$
Ⅱ	272	273	287	287			282	$=69.375$
$\frac{Ⅰ}{4}-\hat{\mu}$	1.375	1.125	−2.375	−2.375			−1.125	
$\frac{Ⅱ}{4}-\hat{\mu}$	−1.375	−1.125	2.375	2.375			1.125	
S	$\frac{121}{8}$	$\frac{81}{8}$	$\frac{361}{8}$	$\frac{361}{8}$			$\frac{81}{8}$	

$$S_A=\frac{2}{8}(Ⅰ_1^2+Ⅱ_1^2)-\frac{1}{8}T^2 \quad \left(T=\sum_1^8 y_i\right)$$

$$=\frac{1}{8}(Ⅰ_1-Ⅱ_1)^2$$

这里$Ⅰ_1$,$Ⅱ_1$分别表示第1列的Ⅰ和Ⅱ.

为了进行方差分析,需要求出误差平方和S_ε. 从(3.8)知

$$S_\varepsilon=\sum_1^8(y_i-\bar{y})^2-(S_A+S_B+S_C+S_D+S_E)$$

$$=\frac{34}{8}$$

方差分析表如下:

方差来源	平方和	自由度	F 值	显著性(α=0.05)
A	121/8	1	7.1	
B	81/8	1	4.8	
C	361/8	1	21.2	*
D	361/8	1	21.2	*
E	81/8	1	4.8	
误　差	34/8	2		

因素 C、D 是显著的，故硫酸浓度是 27% 的好，上海产的比天津产的好. 其他因素的两个水平对收率的影响未显示出差异. 考虑到工时及操作方便，知本例中“最好”生产条件应取 A_1，B_1，C_2，D_2，E_2，即反应温度 50℃，反应时间 1 小时，硫酸浓度 27%，硫酸选择上海产的，操作方式是不搅拌(值得注意的是，这个最好条件并未在 8 次试验中做过).

这种用正交表直接进行分析计算的方法带有普遍性. 用其他正交表时，计算方法类似，从略.

***例 3.4**　拉丁方设计与数据分析.

在农业、生物等领域常使用拉丁方设计，它与采用正交表的正交设计有实质的联系，我们可以认为它是一种特殊的正交设计.

先从一简单情况谈起. 设有 4 种小麦品种 A，B，C，D 种在一块方形田里，要比较产量的高低. 为了消除(或减少)土地纵向和横向肥瘠不匀的影响，以比较品种的优劣，将方田分为 16 等分，种法如下：

行＼列	1	2	3	4
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A

每个品种在每行每列中各出现一次．比较各品种的平均产量就可知道各品种的优劣．

定义 3.2 k 个处理① $1,2,\cdots,k$，安排在 $k\times k$ 方阵内，使每个处理在每行每列均出现一次．这样的一个安排叫做 $k\times k$ 拉丁方设计．

显然，在这种设计中，每个处理重复了 k 次．我们说拉丁方设计实质上是三因素的正交设计，即 $L_{k^2}(k^3)$ 型设计．三个因素是：① 行（每行看成一个区组），共有 k 个水平，用 $1,2,\cdots,k$ 表示．② 列，共有 k 个水平，也用 $1,2,\cdots,k$ 表示．③ 处理，共有 k 个水平．$k\times k$ 拉丁方设计可用下列 $L_{k^2}(k^3)$ 型正交表（表 3.3）来实现．

表 3.3

行	列	处　理	数　据
1	1	λ_1	y_1
1	2	λ_2	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	k	λ_k	y_k
2	1	λ_{k+1}	y_{k+1}
2	2	λ_{k+2}	y_{k+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	k	λ_{2k}	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	1	$\lambda_{(k-1)k+1}$	$y_{(k-1)k+1}$
k	2	$\lambda_{(k-1)k+2}$	$y_{(k-1)k+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	k	λ_{k^2}	y_{k^2}

这种正交表一定有．实际上可取

① “处理”两字作广义的理解，见第三章 § 6.

$$\lambda_1=1,\quad \lambda_2=2,\cdots,\quad \lambda_k=k$$
$$\lambda_{k+1}=k,\quad \lambda_{k+2}=1,\cdots,\quad \lambda_{2k}=k-1$$
$$\lambda_{2k+1}=k-1,\quad \lambda_{2k+2}=k,\quad \lambda_{2k+3}=1,\cdots,\quad \lambda_{3k}=k-2$$
$$\lambda_{(k-1)k+1}=2,\quad \lambda_{(k-1)k+2}=3,\cdots,\quad \lambda_{k^2}=1$$

统计分析实为正交设计情形下的数据分析. 用 $B_{j\lambda}$ 表示表 3.3 中第 j 列水平 λ 对应的数据之和,例如 $B_{11}=y_1+y_2+\cdots+y_k$. 用 T 表示全部数据之和,即 $T=\sum_{1}^{k^2} y_i$. 要比较不同处理的效果,则应比较 k 个平均值:

$$\frac{B_{s\lambda}}{k}\quad (\lambda=1,\cdots,k)$$

为了检验各因素(行、列、处理)是否有影响,有下列方差分析表:

方差来源	平　方　和	自由度	显著性
行	$S_{行}=\frac{1}{k}\sum_{\lambda=1}^{k} B_{1\lambda}^2-\frac{T^2}{k^2}$	$k-1$	
列	$S_{列}=\frac{1}{k}\sum_{\lambda=1}^{k} B_{2\lambda}^2-\frac{T^2}{k^2}$	$k-1$	
处理	$S_{处理}=\frac{1}{k}\sum_{\lambda=1}^{k} B_{3\lambda}^2-\frac{T^2}{k^2}$	$k-1$	
误差	$S_{误}=\sum_{1}^{k^2} y_i^2-\frac{T^2}{k^2}-S_{行}-S_{列}-S_{处理}$	$(k-1)(k-2)$	

* **例 3.5**　希腊拉丁方设计.

除了行、列、品种三个因素外再加一个因素——施肥方式,怎样安排试验才能比较不同品种(或不同施肥方式)的优劣? 著名的安排是所谓希腊拉丁方设计. 设各个因素都是四个水平. 四个品种:A,B,C,D;四种施肥方式:$\alpha,\beta,\gamma,\delta$,可如下安排.

行＼列	1	2	3	4
1	A,α	B,β	C,γ	D,δ
2	B,γ	A,δ	D,α	C,β
3	C,δ	D,γ	A,β	B,α
4	D,β	C,α	B,δ	A,γ

含义是：在每一小块田（共 16 块）上拉丁字母表示种植的小麦品种，希腊字母表示施肥方式. 全部拉丁字母构成 4×4 拉丁方设计，全部希腊字母也构成 4×4 拉丁方设计，而且这两个拉丁方设计有特殊关系：(A,α)，(B,β)，…，(D,δ)各出现一次. 上述设计叫做希腊拉丁方设计.

一般情形：设品种有 k 个（水平）：$A_1,A_2,\cdots,A_k$，施肥方式也有 k 个：$\alpha_1,\alpha_2,\cdots,\alpha_k$. 所谓希腊拉丁方是指 $k\times k$ 方块，每个方块上放上一个 A_i、一个 α_j，使得每行中每个 A_i 恰出现一次，每个 α_j 恰出现一次，每列也是如此，而且(A_i,α_j)在 k^2 个方块中恰出现一次. 这实质上是 $L_{k^2}(k^4)$型设计. 可以证明 $k\neq6$ 时这种设计均存在（$k=10$ 时是 1959 年才发现的），但 $k=6$ 时不存在，这是 1901 年用穷举法得到的结论.

希腊拉丁方设计是一种正交设计，可用正交表来进行数据分析.

更一般的是所谓正交拉丁方设计，它实质上是 $L_{k^2}(k^m)$型正交设计. 可以证明，当 k 是素数或素数幂时，存在 $L_{k^2}(k^{k+1})$型正交表. 有兴趣的读者可去参看[5]和[7].

附：考虑因素间交互作用的正交设计简介

我们将以 $L_8(2^7)$为例说明有关的概念和方法. 先引进一般性定义.

定义 3.3 设 $\Lambda=(\lambda_{ij})_{n\times m}$ 是 $L_n(s_1\times\cdots\times s_m)$型正交表，$1\leqslant j_1\leqslant m$，$1\leqslant j_2\leqslant m$，$1\leqslant k\leqslant m$ 且 j_1,j_2,k 两两不等，称第 k 列是第 j_1 列和第 j_2 列的一个交互列，若对一切 $i_1\neq i_2$（$1\leqslant i_1\leqslant n$，$1\leqslant i_2\leqslant n$），只要 $(\lambda_{i_1j_1},\lambda_{i_1j_2})=(\lambda_{i_2j_1},\lambda_{i_2j_2})$就一定成立 $\lambda_{i_1k}=\lambda_{i_2k}$.

对于前面介绍的 $L_8(2^7)$（标准表），不难看出第 3 列是第 1 列和第 2 列的唯一交互列，第 5 列是第 1 列和第 4 列的唯一交互列，…. 对于这个 $L_8(2^7)$，任何两列都恰有一个交互列，见下表.

$L_8(2^7)$ 标准表的交互列表

列号 \ 交互列号 \ 列号	1	2	3	4	5	6	7
1		3	2	5	4	7	6
2			1	6	7	4	5
3				7	6	5	4
4					1	2	3
5						3	2
6							1
7							

注意，由于第 j_1 列和第 j_2 列的交互列也是第 j_2 列与第 j_1 列的交互列，而且任何两列的交互列只有一个，故在上面的表中只对 $j_1<j_2$ 时给出第 j_1 列和第 j_2 列的交互列是第几列. 例如第 2 列和第 4 列的交互列是第 6 列.（注意，对于 $L_9(3^4)$ 来说，任何两列的交互列是其余两列，对于 $L_{27}(3^{13})$ 和 $L_{16}(2^{15})$ 则有较复杂的交互列表！见[20]. 还有正交表不含有交互列，例如 $L_{12}(2^{11})$.）

下面举例说明考虑某些交互作用时如何进行正交设计.

例 3.6 考察拖拉机在不同作业速度下某些部件对驾驶员耳旁噪声的影响. 拟定的因素及其水平如下表所示，并要考虑交互作用 $A\times B$ 和 $A\times C$ 的影响. 试验指标是耳旁噪声，指标值越小越好.[20]

因素水平表

水平 \ 因素	A（速度）	B（驾驶室）	C（轮胎）	D（风扇）
1	Ⅲ挡	开式	通用加宽	改进型
2	Ⅱ挡	闭式	越　野	普通型

正交设计的步骤是：

(1) 选取合适的正交表．本例中每个因素两个水平，4 个因素加上要考虑的两个“交互作用”共六个“因素”(每个“交互作用”看成一个“因素”)．选用正交表 $L_8(2^7)$ 是合适的，其列数超过 6.

(2) 将字母 A,B,C,D 适当地放在 $L_8(2^7)$ 的某 4 个列的上头(这叫做表头设计)，原则是：避免单个因素与要考虑的“交互作用”混杂，其具体含义是：使得每个因素所在的列不能是要考虑其交互作用的两个因素所在列的交互列．例如，若将 A 放在 $L_8(2^7)$ 标准表的第 1 列上，B 放在第 2 列上，由于要考虑交互作用 $A\times B$，则 C,D 都不能放在第 3 列上(因为第 3 列是第 1 列和第 2 列的交互列)，若将 C 放在第 4 列上，则第 5 列上不能放 D(因为第 5 列是第 1 列与第 4 列的交互列，而交互作用 $A\times C$ 又是我们要考虑的)，D 可以放在第 7 列上(也可以放在第 6 列上)．这样在 $L_8(2^7)$(标准表)的某些列上安排了因素，有些列体现了“交互作用”，此外是空列(未安排因素也不体现要考虑的“交互作用”)．刚才的安排形成“表头设计”：

因素	A	B	$A\times B$	C	$A\times C$		D
列号	1	2	3	4	5	6	7

(3.13)

要注意的是，“表头设计”不是唯一的．将 A,B,C,D 分别放在第 1 列、第 4 列、第 2 列、第 7 列上也可以，此时第 5 列体现交互作用 $A\times B$，第 6 列体现交互作用 $B\times C$.

(3) 写出试验方案

在 $L_8(2^7)$ 上按照上述表头设计(3.13)将因素 A,B,C,D 所在的列(即第 1、2、4、7 列)取出来，把各列中的数字换成对应因素的水平，1 对应水平 1，2 对应水平 2，这就构成了试验方案．内容如下．

拖拉机噪声试验方案表

试验号 \ 因素	(1)A 速 度	(2)B 驾驶室	(4)C 轮 胎	(7)D 风 扇
1	(1)Ⅲ挡	(1)开式	(1)普通加宽	(1)改进型
2	(1)Ⅲ挡	(1)开式	(2)越野	(2)普通型
3	(1)Ⅲ挡	(2)闭式	(1)普通加宽	(3)普通型

续表

试验号 \ 因素	(1)A 速 度	(2)B 驾驶室	(4)C 轮 胎	(7)D 风 扇
4	(1)Ⅲ挡	(2)闭式	(2)越野	(1)改进型
5	(2)Ⅱ挡	(1)开式	(1)普通加宽	(2)普通型
6	(2)Ⅱ挡	(1)开式	(2)越野	(1)改进型
7	(2)Ⅱ挡	(2)闭式	(1)普通加宽	(1)改进型
8	(2)Ⅱ挡	(2)闭式	(2)越野	(2)普通型

以上就是用正交表设计试验方案,即进行正交设计. 下面仍以上述拖拉机噪声试验为例,粗略介绍正交设计下的数据分析方法. 已知按上述试验方案进行试验获得了试验数据 $y_1, y_2, \cdots, y_8$,这里 y_i 是第 i 号试验下驾驶员的耳旁噪声值(单位是分贝,即 dB). 数据及处理方法见下表.

试验结果分析表

试验号 \ 因素	A 1	B 2	$A\times B$ 3	C 4	$A\times C$ 5	6	D 7	y/dB
1	1	1	1	1	1	1	1	92
2	1	1	1	2	2	2	2	98
3	1	2	2	1	1	2	2	94
4	1	2	2	2	2	1	1	97
5	2	1	2	1	2	1	2	94
6	2	1	2	2	1	2	1	93
7	2	2	1	1	2	2	1	86
8	2	2	1	2	1	1	2	91
Ⅰ	381	377	367	366	370	374	368	
Ⅱ	364	368	378	379	375	371	377	
$\frac{1}{4}$Ⅰ	95.25	94.25	91.75	91.50	92.50	93.50	92.00	
$\frac{1}{4}$Ⅱ	91.00	92.00	94.50	94.75	93.75	92.75	94.25	
R	4.25	2.25	2.75	3.25	1.25	0.75	2.25	$R=\left\lvert \frac{\text{Ⅰ}}{4}-\frac{\text{Ⅱ}}{4} \right\rvert$

表中各列的Ⅰ表示“1”对应的数据之和，各列的Ⅱ表示“2”对应的数据之和．例如，第1列的$\text{Ⅰ}=y_1+y_2+y_3+y_4=381$，第1列的$\text{Ⅱ}=y_5+y_6+y_7+y_8=364$；对于放上因素的各列来说，各列的$\frac{1}{4}\text{Ⅰ}$表示该列上的因素的水平1下数据的平均值，各列的$\frac{1}{4}\text{Ⅱ}$表示该列上因素的水平2下的数据的平均值；极差$R=\left|\frac{1}{4}\text{Ⅰ}-\frac{1}{4}\text{Ⅱ}\right|$则表示该列上的因素的两个不同水平对指标（噪声值）带来的差异．显然极差越大，则不同水平带来的差异越大．对于体现“交互作用”的各列来说，可以证明各列对应的极差R正是刻画该列所体现的“交互作用”的大小，R越大表明“交互作用”越大，R太小则表明该“交互作用”可以忽略．（数学证明从略）．总之，按照极差从大到小，将因素（包括“交互作用”）排上次序，以显示其重要程度．在本项试验中，各因素及“交互作用”按极差从大到小的次序是

$$A,\quad C,\quad A\times B,\quad \begin{matrix}B\\D\end{matrix},\quad A\times C$$

$A\times C$所在列的“极差”相当小，可以忽略．C的两个水平中C_1比C_2好，D的两个水平中D_1比D_2好$\left(\text{因为相应的}\frac{1}{4}\text{Ⅰ比}\frac{1}{4}\text{Ⅱ小}\right)$．而$A\times B$所在的第3列对应的“极差”超过因素$B$所在列对应的“极差”，因而要重视$A\times B$，即要研究因素$A$与$B$的水平搭配对指标的影响．为此要研究二元表：

$A \backslash B$	B_1	B_2
A_1	$\frac{92+98}{2}=95$	$\frac{94+97}{2}=95.5$
A_2	$\frac{94+93}{2}=93.5$	$\frac{86+91}{2}=88.5$

这个表用来表示因素A和B的各种水平搭配下试验数据的平均值．例如A取水平1、B取水平1（即A_1B_1）下共有两号试验：第1号和第2号，试验数据的平均值为$\frac{y_1+y_2}{2}=\frac{92+98}{2}=95$，$A$取水平1、$B$取水平2（即$A_1B_2$）下也有两号试验：第3号和第4号，试验数据的平均值为$\frac{y_3+y_4}{2}=\frac{94+97}{2}=95.5$，其

余均类似地算出.

从这个二元表看出，搭配 A_2B_2 最好(因为平均噪声最小).

总之，四个因素的水平的优良组合是 $A_2B_2C_1D_1$，这就找出了优良的工程条件：速度用Ⅱ挡、驾驶室用闭式、轮胎用普通加宽而风扇用改进型. 此条件下驾驶员耳旁噪声最小.

以上是用一个例子粗略说明考虑交互作用时的正交设计及数据分析方法(未介绍方差分析法)，所用方法可以推广到较大的表 $L_{16}(2^{15})$ 和 $L_{32}(2^{31})$ 上去，虽然所遇到的情况比 $L_8(2^7)$ 时要复杂得多；也可推广至 $L_{27}(3^{13})$ 及某些更复杂的表上去，但情况更为复杂. (例如 $L_{27}(3^{13})$ 中任何两列的交互列有两列). 要了解这些复杂情形下的正交设计，请看专著[7]和[20].

我们指出，还有许多灵活使用正交设计的方法，例如拟因子法、拟水平法、并列法、直积法、部分追加法，等等. 可参看[20].

*§4 平衡不完全区组(BIB)设计

这种设计在农业、林业中较多使用，从一个例子讲起.

例 4.1 设有四种小麦品种：A、B、C、D，又有四个区组，每个区组恰有三个单元. 同一区组里土壤条件一样，不同区组可能有较大差异. 问：如何安排品种种植以比较各品种产量的高低？

有一种试验安排是这样的：

Ⅰ	Ⅱ	Ⅲ	Ⅳ	
B	A	C	B	
A	B	A	D	(Ⅰ，Ⅱ，Ⅲ，Ⅳ是四个区组代号)
C	D	D	C	

(把区组内的品种写在该区组的代号下方). 这种安排有如下特点：在同一区组内不能将所有品种都安排下(不完全性)，只安排三个不同的品种. 每个品种出现次数一样多(都是三次)，而且每一对品种出现在同一区组的次数一样多(都是二次). 这种设计叫做平衡不完全区组设计，简称 BIB 设计.

BIB 设计的一般定义是：设有 t 个处理(如品种)，b 个区组，

区组的大小是k(即各个区组一样大,每个区组只能含有 k 个处理). $k<t$,若在每个区组里都安排下 k 个处理使得 b 个区组里的安排有如下性质:每个处理出现次数一样多(叫做处理的重复数),每一对处理出现在同一区组的次数一样多(叫做相遇数),则称这样的安排是 BIB 设计.

在 BIB 设计中,用 r 表示处理的重复数,用 λ 表示相遇数.

定理 4.1 在 BIB 设计中,各参数有下列关系:

$$bk=tr$$

$$kr=r+\lambda(t-1)$$

证明:第一式是显然的,现证第二式. 任固定一个处理 A_i 及处理 $A_j(i\neq j)$,$\{A_i, A_j\}$共在 λ 个区组中出现. 含有 A_i 的区组共 r 个,这样的每个区组包含 A_i 以外的处理有 $k-1$ 个,因而这样的 r 个区组总共有 $r(k-1)$个单元. 另一方面,A_i 以外的处理共 $t-1$ 个,每个出现 λ 次,故 $\lambda(t-1)$个填满 $r(k-1)$个单元,所以 $r(k-1)=\lambda(t-1)$. 证毕.

由此可见,在 BIB 设计中只要 t,b,k 确定了,则 r 与 λ 就确定了.

在 BIB 设计中,通常写出四个参数(t,b,k,r).

例 4.2 BIB 设计(3,3,2,2).

区组 / 处理	Ⅰ	Ⅱ	Ⅲ
A	×	×	
B	×		×
C		×	×

(含义是区组Ⅰ包含处理 A,B;区组Ⅱ包含处理 A,C;区组Ⅲ包含处理 B,C).

例 4.3 BIB 设计(4,6,2,3).

	I	II	III	IV	V	VI
A	×				×	×
B	×	×		×		
C		×	×		×	
D			×	×		×

给定参数 t,b,k,怎样进行 BIB 设计？现代已有大量研究. 构造理论属于组合数学的范围,读者可从[5]或[7]中查到试验工作中用到的设计. 要注意的是,并不是对所有的 t,b,k,BIB 设计都存在.

BIB 设计不是正交设计. 我们用 $1,2,\cdots,t$ 表示 t 个处理,用矩阵$(n_{ij})_{t\times b}$表示 BIB 设计,这里

$$n_{ij}=\begin{cases}1, & \text{当处理 } i \text{ 在区组 } j \text{ 中出现}\\ 0, & \text{否则}\end{cases}$$

不难看出

$$\sum_{j=1}^{b} n_{ij}=r \quad (\text{处理的重复数})$$

$$\sum_{i=1}^{t} n_{ij}=k \quad (\text{区组大小})$$

$$\sum_{l=1}^{b} n_{il}n_{jl}=\begin{cases}\lambda, & (\text{相遇数}) \quad \text{当 } i\neq j\\ r, & \text{当 } i=j\end{cases} \tag{4.1}$$

现在来介绍 BIB 设计下的统计分析.

令 $D=\{(i,j):n_{ij}=1\}$,当 $i,j\in D$ 时,用 y_{ij} 表示区组 j 里处理 i 对应的指标观测值(例如产量). 我们假定有下列可加模型

$$y_{ij}=\mu+\alpha_i+\beta_j+e_{ij} \quad ((i,j)\in D)$$

其中 α_i 是处理效应,β_j 是区组效应,适合约束条件:

$$\sum_{i=1}^{t}\alpha_i=0,\quad \sum_{j=1}^{b}\beta_j=0$$

为了找出 μ,α_i,β_j 的最小二乘估计 $\hat{\mu},\hat{\alpha}_i,\hat{\beta}_j$,我们采用拉格朗

日乘子法,令

$$Q=\sum_{(i,j)\in D}(y_{ij}-\mu-\alpha_i-\beta_j)^2+2\lambda_1\sum_1^t\alpha_i+2\lambda_2\sum_1^b\beta_j$$

解方程组:

$$\begin{cases}\dfrac{\partial Q}{\partial\mu}=0 & (4.2)\\ \dfrac{\partial Q}{\partial\alpha_i}=0\quad(i=1,2,\cdots,t) & (4.3)\\ \dfrac{\partial Q}{\partial\beta_j}=0\quad(j=1,2,\cdots,b) & (4.4)\\ \sum_1^t\alpha_i=\sum_1^b\beta_j=0 & (4.5)\end{cases}$$

从(4.2)知

$$\sum_{(i,j)\in D}y_{ij}-N\hat{\mu}=0$$

(N 等于 D 中元素个数,即数据$\{y_{ij}\}$的个数),故

$$\hat{\mu}=\frac{1}{N}\sum_{(i,j)\in D}y_{ij}=\bar{y}_{..}\quad(\text{全部数据的平均值})$$

从(4.3)和(4.4)知

$$y_{i.}-r\hat{\mu}-r\hat{\alpha}_i-\sum_{j=1}^b n_{ij}\hat{\beta}_j=\lambda_1 \tag{4.6}$$

$$y_{.j}-k\hat{\mu}-k\hat{\beta}_j-\sum_{i=1}^t n_{ij}\hat{\alpha}_i=\lambda_2 \tag{4.7}$$

这里 $y_{i.}=\sum_{j:(i,j)\in D}y_{ij}\quad y_{.j}=\sum_{i:(i,j)\in D}y_{ij}$

(4.6)式两端对 i 求和得

$$\lambda_1=r\bar{y}_{..}-r\hat{\mu} \tag{4.8}$$

(4.7)式两端对 j 求和得

$$\lambda_2=k\bar{y}_{..}-k\hat{\mu} \tag{4.9}$$

从(4.6)及(4.8)知

$$ky_{i.}-rk\hat{\mu}-rk\hat{\alpha}_i-k\sum_{j=1}^b n_{ij}\hat{\beta}_j=kr\bar{y}_{..}-rk\hat{\mu} \tag{4.10}$$

从(4.7)及(4.9)知

$$\sum_{j=1}^{b} n_{ij} y_{\cdot j} - rk\hat{\mu} - k\sum_{j=1}^{b} n_{ij}\hat{\beta}_j - \sum_{l=1}^{t}\left(\sum_{j=1}^{b} n_{lj} n_{ij}\right)\hat{\alpha}_l$$

$$= kr\bar{y}_{\cdot\cdot} - rk\hat{\mu} \tag{4.11}$$

从(4.10)的两端分别减去(4.11)式的两端,并注意(4.1)得到

$$\hat{\alpha}_i = \frac{1}{\lambda t}\left(ky_{i\cdot} - \sum_{j=1}^{b} n_{ij} y_{\cdot j}\right) \quad (i=1,\cdots,t) \tag{4.12}$$

从(4.7)和(4.9)得

$$\hat{\beta}_j = \frac{1}{k}\left(y_{\cdot j} - \sum_{i=1}^{t} n_{ij}\hat{\alpha}_i\right) - \bar{y}_{\cdot\cdot} \quad (j=1,\cdots,b)$$

为了检验不同处理对于指标是否有不同影响,应检验下列假设

$$H_0: \alpha_1 = \cdots = \alpha_t$$

采用统计量

$$F = \frac{S_{\alpha t}/(t-1)}{S_e/(bk-t-b+1)} \tag{4.13}$$

其中

$$S_{\alpha t} \overset{\mathrm{d}}{=\!=} \frac{k}{\lambda t}\sum_{i=1}^{t}\left(y_{i\cdot} - \frac{1}{k}\sum_{j=1}^{b} n_{ij} y_{\cdot j}\right)^2$$

$$= \frac{\lambda t}{k}\sum_{1}^{t}(\hat{\alpha}_i)^2 \tag{4.14}$$

$$S_e \overset{\mathrm{d}}{=\!=} \sum_{(i,j)\in D}(y_{ij} - \bar{y}_{\cdot\cdot})^2 - S_{\alpha t} - k\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 \tag{4.15}$$

$$\bar{y}_{\cdot j} \overset{\mathrm{d}}{=\!=} \frac{1}{k} y_{\cdot j}$$

可以证明(参看[7]),在 H_0 下,F 服从自由度为 $t-1$, $bk-t-b+1$ 的 F 分布. 给定显著性水平 α,查 F 分布表得临界值 F_0. 当 $F>F_0$ 时拒绝 H_0,即认为不同处理对指标的影响有差异.

例 4.4 某化工产品生产过程的反应时间依赖于所用催化剂的种类,希望通过试验比较 4 种催化剂的作用. 考虑到原料的差异可能干扰催化剂的比较,决定将原料批取作区组因素. 但因只有 4 批原料,而每批原料只够 3 种催化剂做试验,故只能使用不完全区组设计.

在此试验中,要比较的处理数 $t=4$,四种催化剂分别用 1,2,3,4 表示;区组数:$b=4$,四批原料分别用 1,2,3,4 表示;区组大小 $k=3$. 此时的一种 BIB 设计为(4,4,3,3).

处理＼区组	1	2	3	4
1	×	×		×
2		×	×	×
3	×	×	×	
4	×		×	×

(含义是区组 1 包含处理 1,2,4,即第一批原料分别用第 1,2,4 种催化剂进行试验,其余类推).

试验结果见下表(单位:分)

处理＼区组	1	2	3	4	和 $y_{i\cdot}$
1	73	74		71	218
2		75	67	72	214
3	73	75	68		216
4	75		72	75	222
和 $y_{\cdot j}$	221	224	207	218	$y_{\cdot\cdot}=870$

设四种催化剂的效应分别为 $\alpha_1,\alpha_2,\alpha_3,\alpha_4$,从(4.12)知

$$\hat{\alpha}_i=\frac{1}{8}\left(3y_{i\cdot}-\sum_{j=1}^{4}n_{ij}y_{\cdot j}\right)\quad(i=1,2,3,4)$$

于是

$$\hat{\alpha}_1=\frac{1}{8}[3y_{1\cdot}-(y_{\cdot1}+y_{\cdot2}+y_{\cdot4})]=-\frac{9}{8}$$

$$\hat{\alpha}_2=\frac{1}{8}[3y_{2\cdot}-(y_{\cdot2}+y_{\cdot3}+y_{\cdot4})]=-\frac{7}{8}$$

$$\hat{\alpha}_3=\frac{1}{8}[3y_{3\cdot}-(y_{\cdot1}+y_{\cdot2}+y_{\cdot3})]=-\frac{4}{8}$$

$$\hat{\alpha}_4=\frac{20}{8}$$

可见使用第 1 种催化剂反应时间最短. 为了检验假设:

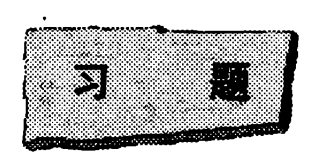
$$H_0:\alpha_1=\alpha_2=\alpha_3=\alpha_4$$

可采用统计量(4.13). 在 H_0 下统计量 F 服从自由度为 3,5 的 F 分布. 从(4.14)知 $S_\alpha=22.75$,从(4.15)知 $S_e=3.25$. 于是

$$F=\frac{22.75/3}{3.25/5}=11.67$$

若显著性水平 $\alpha=0.05$,查 F 分布表知临界值 $F_0=5.41$,于是应拒绝 H_0.

故催化剂的种类对反应时间有显著性影响,采用第 1 种催化剂反应时间确实最短.



1. 一工厂有设备 A_1,A_2,A_3 生产某种产品,连续五日记录了这三台设备的每日次品率(百分数).

	日次品率				
A_1	4.1	4.8	4.1	4.9	5.7
A_2	6.1	5.7	5.4	7.2	6.4
A_3	4.5	4.8	4.8	5.1	5.6

就上列数据分析三台设备有无显著性差异?(显著性水平 $\alpha=0.05$)

2. 为了试制一种化工产品. 在三种不同温度、四种不同压力下做试验,

每一水平组合重复两次，得产品的收率数据如下（%）：

温度 \ 压力	1	2	3	4
1	(52，57)	(42，45)	(41，45)	(48，45)
2	(50，52)	(47，45)	(47，48)	(53，50)
3	(63，58)	(54，59)	(57，60)	(53，59)

问：温度、压力及温度与压力的交互作用中哪些对收率有显著影响？（显著性水平 $\alpha=0.05$）

3. 在实际进行方差分析的计算时，为了计算的方便，常常对观测数据先作一线性变换. 以单因素方差分析为例，令 $x_{ij}=(y_{ij}-c)/d$，其中 c,d 是固定常数（$d\neq 0$），试用 x_{ij} 表示出单因素的方差分析表.

4. 在单因素方差分析中，若各水平下的试验重复数不相等，即有观测值 y_{ij} ($i=1,2,\cdots,s$ $j=1,2,\cdots,r_i$，这里 s 是水平个数，$s\geqslant 2$，r_i 是水平 i 的重复次数，$r_i\geqslant 1$). 这里

$$y_{ij}=\mu_i+e_{ij}$$

$\{e_{ij}\}$ 相互独立同分布，$e_{ij}\sim N(0,\sigma^2)$ (σ^2 未知).

问：在此情形下如何检验 $H_0:\mu_1=\mu_2=\cdots=\mu_s$.

5. 在 $k\times k$ 拉丁方设计中，设第 i 行第 j 列施加处理 l 得到的观测值为 $y_{ij(l)}$，数学模型为

$$y_{ij(l)}=\mu+\alpha_i+\beta_j+\gamma_l+e_{ij}\ ,\quad i,j=1,2,\cdots,k$$

其中 $\{e_{ij}\}$ 相互独立同分布，$e_{ij}\sim N(0,\sigma^2)$ (σ^2 未知) 且

$$\sum_{i=1}^{k}\alpha_i=\sum_{j=1}^{k}\beta_j=\sum_{l=1}^{k}\gamma_l=0$$

（注意给定拉丁方设计后，l 由 i,j 唯一确定，$l=l(i,j)$）. 令

$$\bar{y}=\frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}y_{ij(l)}\quad \text{（全部数据的平均值）}$$

$$\bar{y}_{i\cdot}=\frac{1}{k}\sum_{j=1}^{k}y_{ij(l)}\quad \text{（第 } i \text{ 行全部数据的平均值）}$$

$$\bar{y}_{\cdot j}=\frac{1}{k}\sum_{i=1}^{k}y_{ij(l)}\quad \text{（第 } j \text{ 列全部数据的平均值）}$$

$$\bar{y}_{(l)}=\frac{\text{处理为 } l \text{ 的全部数据之和}}{k}$$

试证明下列恒等式

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(y_{ij(l)}-\bar{y})^2=k\sum_{i=1}^{k}(\bar{y}_{i\cdot}-\bar{y})^2+k\sum_{j=1}^{k}(\bar{y}_{\cdot j}-\bar{y})^2$$
$$+k\sum_{l=1}^{k}(\bar{y}_{(l)}-\bar{y})^2+\sum_{i=1}^{k}\sum_{j=1}^{k}(y_{ij(l)}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}$$
$$-\bar{y}_{(l)}+2\bar{y})^2$$

并用上述这些平方和列出拉丁方的方差分析表.

6. 在冲天炉降低焦比的试验中,考察的指标是铁水温度. 试验目的是提高铁水温度,降低焦炭的用量. 考察底焦高度(m)、每批焦比、风压(mmHg)与风吡(%)这四个因素对铁水温度的影响. 选出的四个因素都考察三个水平,即

水平 \ 因素	每批焦比	风压	风吡	底焦高度
1 水平	1∶17	180	1.1	1.2
2 水平	1∶19	240	0.78	1.5
3 水平	1∶15	210	0.89	1.4

将这四个因素依次放在 $L_9(3^4)$ 的第 1,2,3,4 列上,就可得到试验计划. 现将试验计划及按计划试验后得到的数据(铁水温度(℃))列表如下:

试号 \ 列号	每批焦比 1		风压 2		风吡 3		底焦高度 4		铁水温度
1	1	1∶17	1	180	1	1.1	1	1.2	1 369
2	1	1∶17	2	240	2	0.78	2	1.5	1 393
3	1	1∶17	3	210	3	0.89	3	1.4	1 385
4	2	1∶19	1	180	2	0.78	3	1.4	1 394
5	2	1∶19	2	240	3	0.89	1	1.2	1 394
6	2	1∶19	3	210	1	1.1	2	1.5	1 378
7	3	1∶15	1	180	3	0.89	2	1.5	1 391
8	3	1∶15	2	240	1	1.1	3	1.4	1 411
9	3	1∶15	3	210	2	0.78	1	1.2	1 392

试对上述试验数据进行分析.

7. 为了考察四个因素 A,B,C,D 对指标 y 的影响. 每个因素取两个水平(分别用 $A_1,A_2,B_1,B_2,C_1,C_2,D_1,D_2$ 表示),将 A,B,C,D 放在正交表 $L_8(2^7)$ 的第 1,2,4,5 列上,就可得到试验计划. 每个试验条件下做两次试验. 现将试验计划及试验得到的数据列在下表:

列号 行号	A 1	B 2	3	C 4	D 5	6	7	y	
1	1	1		1	1			7	9
2	1	1		2	2			10	11
3	1	2		1	2			9	10
4	1	2		2	1			12	13
5	2	1		1	1			9	10
6	2	1		2	2			12	11
7	2	2		1	2			11	11
8	2	2		2	1			15	16

试对数据进行统计分析(希望提高指标值).

8. 设 $\Lambda=(\lambda_{ij})$ 是 $L_n(2^m)$ 型正交表且 $n\neq 2$, $m\neq 1$,试证明 n 必是 4 的倍数.

9. 设 $\Lambda=(\lambda_{ij})$ 是 $L_n(2^m)$ 型正交表. 令

$$\tilde{\lambda}_{ij}=\begin{cases}2, & \text{当 } \lambda_{ij}=1\\ 1, & \text{当 } \lambda_{ij}=2\end{cases}$$

$$\tilde{\Lambda}=(\tilde{\lambda}_{ij})$$

$$M=\left(\begin{array}{c:c} \begin{matrix}1\\ \vdots\\ 1\end{matrix}\quad \Lambda & \Lambda \\ \hdashline \begin{matrix}2\\ \vdots\\ 2\end{matrix}\quad \Lambda & \tilde{\Lambda}\end{array}\right)$$

试证明:(1) M 是 $L_{2n}(2^{2m+1})$ 型正交表;

(2) 试构造出一张 $L_{16}(2^{15})$ 型正交表.

10. 设 $\Lambda=(\lambda_{ij})$ 是 $L_n(3^m)$ 型正交表，

$$\tilde{\lambda}_{ij}=\begin{cases}2, & 当\ \lambda_{ij}=1\\ 3, & 当\ \lambda_{ij}=2\\ 1, & 当\ \lambda_{ij}=3\end{cases}$$

$$\hat{\lambda}_{ij}=\begin{cases}3, & 当\ \lambda_{ij}=1\\ 1, & 当\ \lambda_{ij}=2 \quad (一切\ i,j)\\ 2, & 当\ \lambda_{ij}=3\end{cases}$$

$$\tilde{\Lambda}=(\tilde{\lambda}_{ij}),\ \hat{\Lambda}=(\hat{\lambda}_{ij})$$

$$M=\begin{pmatrix} 1 & & & \\ \vdots & \Lambda & \Lambda & \Lambda \\ 1 & & & \\ \hdashline 2 & & & \\ \vdots & \Lambda & \tilde{\Lambda} & \hat{\Lambda} \\ 2 & & & \\ \hdashline 3 & & & \\ \vdots & \Lambda & \hat{\Lambda} & \tilde{\Lambda} \\ 3 & & & \end{pmatrix}$$

(1) 试证明 M 是 $L_{3n}(3^{3m+1})$ 型正交表；

(2) 试构造出一张 $L_{27}(3^{13})$ 型正交表.

第六章 序贯分析初步

§1 序贯方法的重要性与两个要素

大家知道,统计学里最基本的概念是总体与样本. 总体常用随机变量(或随机向量)来刻画,它的特性是用分布函数来描述,样本则是由总体中抽取的(或观测到的)若干个个体(样单元)组成,它是由一组随机变量(向量)来刻画,样本中所含的样单元个数叫做样本量. 统计学的基本任务是通过样本来推测或掌握总体的特性. 一个很基本的问题是:样本量要取多大? 当然样本量越大,对总体的了解就越多,因此对总体作出的推断也就越可靠. 但是抽样(或进行观测)是需要费用的,抽样量越大,耗费就越大. 合理的提法是,在保证所得结论有足够可靠性的前提下,应使得抽样量越小越好. 通常的统计方法都是在抽样之前预先给定抽样量的大小,实践表明,这种固定样本量的方法能解决很多问题,但在有些情况下使用这种方法导致不必要的浪费. 还存在这样的情况:使用固定样本量方法(无论样本量多大),根本不能解决问题. 请看下列例子.

例 1.1(单式抽样方案) 设有一批产品(共 N 件,N 很大)需经验收检验. 每个产品经过检验可判定为合格还是不合格. 最简单的验收方案是选定两个整数 $n>c\geqslant 0$,从该批产品中随机抽取 n 件,如果这 n 件中所含的不合格品件数 $d>c$,则拒收该批产品;若 $d\leqslant c$,则接收该批产品. 很明显,这个验收方案的样本量(抽样量)是固定的整数 n,但是在抽样的过程中,如果未抽到 n 件就已抽到了 $c+1$ 个不合格品,就没有必要再往下抽样. 换句话说,在有些

情况下事先固定样本量要造成浪费. 这就启示我们:应根据抽样过程中出现的情况来决定抽样量.

例 1.2 研究一枚不匀称的硬币. 我们把有币值的一面叫正面,另一面叫反面. 在桌上任意抛掷一次,问:正面朝上的概率是否大于$\frac{1}{2}$? 我们可以用随机变量 X 刻画这枚硬币. 当正面朝上时,$X=1$;反面朝上时 $X=0$,记 $p=P(X=1)$. 当然 $0<p<1$. 上述问题可化为检验零假设 $H_1: 0<p<\frac{1}{2}$ $\left(\text{备择假设是 } H_2: \frac{1}{2}<p<1\right)$. 若拒绝 H_1,则表明正面朝上的概率大于$\frac{1}{2}$;若接受 H_1,则表明正面朝上的概率小于$\frac{1}{2}$. 在进行检验时,每"抽一个样"就是将硬币任意抛掷一次. 给定小正数 $\alpha\left(0<\alpha<\frac{1}{2}\right)$,问:是否有固定样本量的检验法使得犯两类错误的概率都一致小于 α 呢? 可以证明,这个问题的答案是否定的,也就是说,无论预先固定的样本量有多大,也无济于事.

实际上,设 $X_1,X_2,\cdots,X_n$ 是 X 的样本,W 是样本量等于 n 的否定域且

$$\sup_{0<p<\frac{1}{2}} P_p((X_1,\cdots,X_n)\in W)\leqslant \alpha$$

(这里 $P_p(A)$表示参数是 p 时事件 A 的概率).

因为

$$P_p((X_1,\cdots,X_n)\in W) = \sum_{(x_1,\cdots,x_n)\in W} p^{\sum_1^n x_i}(1-p)^{n-\sum_1^n x_i}$$

是 p 的连续函数(注意,$(X_1,\cdots,X_n)$的可能值共有 2^n 个),于是

$$\lim_{p\to\frac{1}{2}+0} P_p((X_1,\cdots,X_n)\in W^c) = \lim_{p\to\frac{1}{2}-0} P_p((X_1,\cdots,X_n)\in W^c) \geqslant 1-\alpha > \frac{1}{2} > \alpha$$

这里 W^c 是 W 的余集. 故采用否定域 W 犯第二类错误的概率不能一致地不大于 α.

例 1.3 设 $X \sim N(\mu, \sigma^2)$，μ、σ 都未知. 给定 $\gamma \in (0,1)$ 及 $L > 0$，问：μ 是否有固定样本量的置信区间 $[\varphi_1(X_1, \cdots, X_n), \varphi_2(X_1, \cdots, X_n)]$ 使得置信水平是 γ，区间宽度不超过 L？可以证明，这个问题的答案是否定的. 也就是说，无论固定的样本量多大，不存在固定宽度的置信区间(详见 § 3).

上述三个例子是有代表性的，表明了固定样本量方法的局限性. 第一个例子表明固定样本量方法有时效率不高，第二、三个例子表明固定样本量方法有时无能为力. 统计学发展史上的一个重要里程碑就是序贯方法的出现. 人们在 20 世纪 40 年代普遍认识到样本量不必预先固定，可以根据抽样(或观测)过程出现的情况来决定何时停止抽样(或观测)，也就是说样本量是一个随机变量，这样得到的样本是一个一个地逐次得到的，叫做序贯样本. 可以证明，在例 1.2 中存在基于序贯样本的检验方法，其两类错误概率均一致不超过 α，在例 1.3 中存在基于序贯样本的固定宽度的置信区间.

序贯分析(或者说统计中的序贯方法)正是研究如何得到和利用序贯样本进行统计推断(或选择)的统计学分支. 在假设检验、参数估计及更一般的统计判决问题里，序贯分析方法一般有两个组成部分(两个要素)：停止法则①与判决法则. 停止法则告诉我们，在对总体进行逐次观测(或抽样)的过程中何时停止下来；判决法则则告诉我们，根据停止时得到的全部数据(序贯样本)对总体应如何作出推断或选择(接受或拒绝一个假设，估计

① 有些统计问题里(见 § 4)不明确提出停止法则，允许抽样(或观测)无限进行下去，而每次(或每步)如何抽样(或观测)却需要根据在此之前得到的全部数据才能确定. 这种递推方法也属于序贯分析的范围. 我们可以这样设想，虽然递推过程可以无限进行下去，在实际使用时也是进行到一定阶段(例如估计参数时误差足够小)就停下来，因而还是有一个停止法则在起作用，不过未明确提出罢了.

参数等等).

数学上如何描述停止法则与判别法则呢?停止法则的定义如下:设 $X_1, X_2, \cdots$ 是独立同分布的随机变量列(X_i 与总体 X 有相同的分布).称随机变量 τ 是停止法则,若 τ 只取非负整数值(但可取值 ∞),而且 $\tau \equiv 0$,或 $\tau \geqslant 1$,对一切 $n \geqslant 1$,存在集合 B_n(Borel 集)使得

$$\{\tau \leqslant n\} = \{(X_1, \cdots, X_n) \in B_n\} \tag{1.1}$$

(1.1)式的直观意义是:τ 是否大于 n 仅由观测值 $X_1, \cdots, X_n$ 来确定,而与尚未观测的 $X_{n+1}, X_{n+2}, \cdots$ 无关.这表明停止法则乃是不依赖于将来的随机变量.当 $\tau \equiv 0$ 时,表示不进行任何观测(或抽样).

例 1.4 设 $X_1, X_2, \cdots$ 是随机变量列,B 是直线上的集合(Borel 集),

$$\tau_1 = \inf\{n : n \geqslant 1,\ X_n \in B\}$$

(当不存在 n 使得 $X_n \in B$ 时规定 $\tau_1 = \infty$),

$$\tau_2 = \inf\{n : n \geqslant 1,\ X_{n+1} \in B\}$$

(当不存在 n 使得 $X_{n+1} \in B$ 时规定 $\tau_2 = \infty$),则对一切 $n \geqslant 1$,

$$\{\tau_1 \leqslant n\} = \bigcup_{i=1}^{n} \{X_i \in B\}, \qquad \{\tau_2 \leqslant n\} = \bigcup_{i=1}^{n+1} \{X_i \in B\}$$

可见 τ_1 是停止法则,τ_2 一般不是停止法则.

称停止法则 τ 是封闭的,若它是有限的,即

$$P(\tau < \infty) = 1$$

我们最关心封闭的停止法则.不过有时不封闭的(叫做开放的)停止法则也是有实际意义的.

显然,若 $\tau \equiv n$,则 τ 也是停止法则,故固定样本量情形是一种特殊情形.

停止法则也称停止时间(简称停时).

怎样描述判决法则呢?最一般情形是给定某个非空集合 A(叫做行动空间),A 中的元素叫做"行动".所谓判决法则,乃是

取值于空间 A 的映射 $d=d(X_1,\cdots,X_\tau)$. 换句话说,判决法则 d 告诉我们:在得到序贯样本 $(X_1,\cdots,X_\tau)$ 后应选择什么样的"行动".

停止法则 τ 与判决法则 d 合在一起构成序贯方法 Δ,记为 $\Delta=(\tau,d)$.

读者在这里可能感到停止法则和判决法则的定义过于抽象,但通过后面 §2 的学习就不难理解和掌握这两个概念.

"假设检验"乃是一种两判决问题,即行动空间 A 由两个元素组成,$A=\{a_1,a_2\}$,a_1 理解为接受假设 H_1,a_2 理解为接受假设 H_2 (H_2 是 H_1 的备择假设). §2 中介绍序贯概率比检验,这是最重要的序贯检验法. §3 中介绍序贯估计和随机逼近. 我们的介绍是入门性质的,更全面、更系统的论述要用到很多测度论和现代概率论知识,读者对此如有兴趣可参阅[8].

§2 序贯概率比检验(SPRT)

最基本的序贯方法是序贯分析奠基人 A. Wald 于 1943～1945 年间提出的序贯概率比检验. 这是适应第二次世界大战期间美国军火生产中质量检验工作的需要而创造出来的.

设 X 的分布密度(或概率函数)是 $f_1(x)$ 或 $f_2(x)$,到底是哪一个? 不知道. 我们来检验假设 H_1:f_1 是真正的分布密度(概率函数). 对立假设是 H_2:f_2 是真正的分布密度(概率函数).

对 X 进行观测得到 $X_1,X_2,\cdots$,它们是独立同分布的,共同的分布密度(或概率函数)是 f_1 或 f_2.

为了检验 H_1,考虑似然比统计量:

$$\lambda_n=\prod_{i=1}^{n} f_2(X_i)\Big/\prod_{i=1}^{n} f_1(X_i)\quad (n\geqslant 1)\tag{2.1}$$

在这里我们采用通常的习惯:随机变量与随机变量的观测值在记号上不加区别. $X_1,\cdots,X_n$ 既代表 n 个随机变量又代表这 n 个随

机变量的观测值(样本值).

如果得到了 n 个观测值 $X_1,\cdots,X_n$,根据熟知的 N－P 引理,最好的检验法是似然比检验,即找出临界值 C,当 $\lambda_n>C$ 时拒绝 H_1;当 $\lambda_n\leqslant C$ 时接受 H_1. A. Wald 对此方法进行了重大改进.他的基本思想是:当 λ_n 很大时拒绝 H_1,当 λ_n 很小时接受 H_1,当 λ_n 不太大、也不太小时就不忙做结论而再观测(抽样)一次,往下研究 λ_{n+1},直到某一步似然比足够大或足够小为止. 确切地说,给定两个常数 $A,B(0<A<1<B<\infty)$,Wald 提出的检验法的实施步骤是:设 X_1 是第一次观测得到的值,计算似然比 λ_1(见(2.1)),如果

$$\lambda_1\geqslant B$$

则停止观测并拒绝假设 H_1;如果

$$\lambda_1\leqslant A$$

则停止观测并接受假设 H_1;如果

$$A<\lambda_1<B$$

则进行第二次观测,得观测值 X_2,计算似然比 λ_2(见(2.1)),如果

$$\lambda_2\geqslant B$$

则停止观测并拒绝 H_1;如果

$$\lambda_2\leqslant A$$

则停止观测并接受 H_1;如果

$$A<\lambda_2<B$$

则进行第三次观测,得观测值 X_3……一般地,如果进行了 $n-1$ 个观测不能做出停止观测并拒绝或接受 H_1 的决定,则进行第 n 次观测,得观测值 X_n 并计算似然比 λ_n. 如果 $\lambda_n\geqslant B$,则停止观测并拒绝 H_1;如果 $\lambda_n\leqslant A$,则停止观测并接受 H_1;如果 $A<\lambda_n<B$,则进行第 $n+1$ 次观测,得观测值 X_{n+1}……

定义 2.1 上述检验法叫做序贯概率比检验(SPRT),简记为 $S(A,B)$. 它的停止法则是

$$\tau^*=\inf\{n:n\geqslant 1,\ \lambda_n\overline{\in}(A,B)\}\tag{2.2}$$

（若对一切 $n\geqslant 1$，$\lambda_n\in(A,B)$，则规定 $\tau^*=\infty$），判决法则是

$$d^*=\begin{cases}\text{拒绝 } H_1, & \text{当 } \lambda_{\tau^*}\geqslant B\\ \text{接受 } H_1, & \text{当 } \lambda_{\tau^*}\leqslant A\end{cases}\tag{2.3}$$

从 $S(A,B)$ 的定义看出，最要紧的是找出停止法则的具体表达式.

例 2.1 伯努利分布情形.

设 X 取值 0 或 1，$P(X=1)=p=1-P(X=0)$，$0<p<1$，故 X 的概率函数 $f(x,p)=P(X=x)=p^x(1-p)^{1-x}(x=0,1)$，考虑检验问题：

$$H_1:\ p=p_1\leftrightarrow H_2:\ p=p_2$$

这里 $0<p_1<p_2<1$，p_1，p_2 是已知的. 设 $x_1,x_2,\cdots$ 是 X 的独立观测序列，从(2.1)知

$$\lambda_n=\prod_{i=1}^{n} f(x_i,p_2)\Big/\prod_{i=1}^{n} f(x_i,p_1)$$

$$=\left(\frac{p_2}{p_1}\right)^{s_n}\left(\frac{1-p_2}{1-p_1}\right)^{n-s_n}$$

这里 $s_n=\sum\limits_{1}^{n} x_i$，于是 $\ln\lambda_n=s_n\ln\dfrac{p_2}{p_1}+(n-s_n)\ln\left(\dfrac{1-p_2}{1-p_1}\right)$. 给定 $0<A<1<B$，令

$$c=-\frac{\ln\dfrac{1-p_2}{1-p_1}}{\ln\dfrac{p_2}{p_1}-\ln\dfrac{1-p_2}{1-p_1}}$$

$$d_1=\frac{\ln B}{\ln\dfrac{p_2}{p_1}-\ln\dfrac{1-p_2}{1-p_1}}$$

$$d_2=\frac{\ln A}{\ln\dfrac{p_2}{p_1}-\ln\dfrac{1-p_2}{1-p_1}}$$

$R_n = cn + d_1$, $A_n = cn + d_2$

易知 $c>0$, $d_1>0$, $d_2<0$,不难看出 $\lambda_n \geqslant B$ 的充要条件是 $s_n \geqslant R_n$;$\lambda_n \leqslant A$ 的充要条件是 $s_n \leqslant A_n$,故 SPRT 的停止法则是

$$\tau^* = \inf\{n: n \geqslant 1,\ s_n \overline{\in} (A_n, R_n)\}$$

我们可以用图来表示这个检验方案. 图 17 中横坐标表示抽样量,纵坐标是样本和 s_n 的值. 当点(n,s_n)落在两条平行直线之间时继续抽样(观测),否则停止抽样(观测).

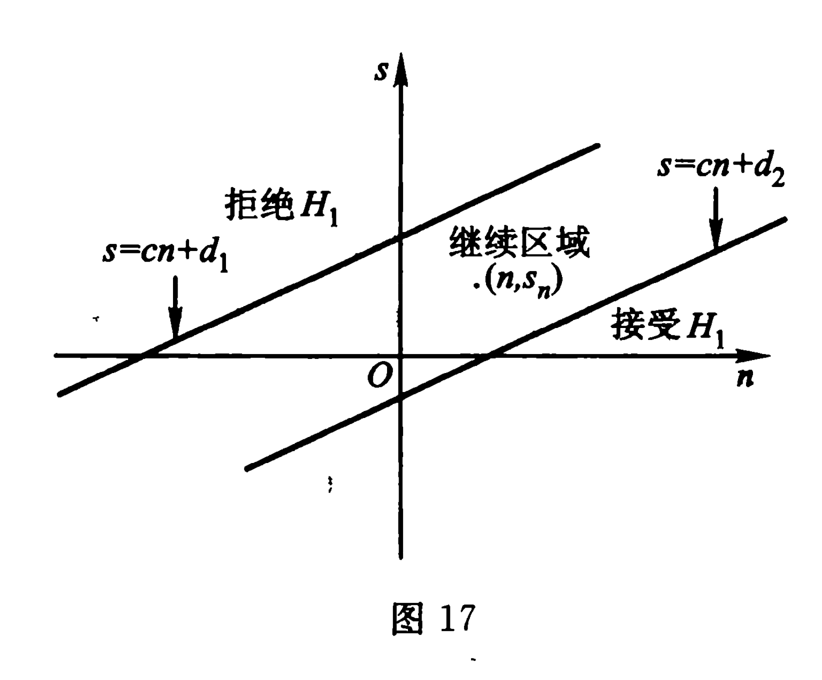


图 17

例 2.2 正态分布情形.

设 $X \sim N(\theta,1)$,$\theta \in (-\infty,\infty)$,分布密度是

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

考虑检验问题:

$$H_1: \theta = \theta_1 \leftrightarrow H_2: \theta = \theta_2$$

这里 $\theta_1 < \theta_2$. 设 $x_1, x_2, \cdots$是 X 的独立观测序列. 从(2.1)知

$$\lambda_n = \prod_{i=1}^{n} f(x_i,\theta_2) \Big/ \prod_{i=1}^{n} f(x_i,\theta_1)$$

$$= e^{-\frac{1}{2}\sum_1^n (x_i-\theta_2)^2} \Big/ e^{-\frac{1}{2}\sum_1^n (x_i-\theta_1)^2}$$

于是

$$\ln \lambda_n = (\theta_2 - \theta_1) s_n + \frac{n}{2}(\theta_1^2 - \theta_2^2),$$

这里 $s_n = \sum_{1}^{n} x_i$. 不难推知，$S(A,B)$的停止法则是

$$\tau^* = \inf\{n: n \geqslant 1,\ s_n \overline{\in} (A_n, R_n)\}$$，其中

$$A_n = \frac{\theta_1 + \theta_2}{2} n + \frac{\ln A}{\theta_2 - \theta_1}$$

$$R_n = \frac{\theta_1 + \theta_2}{2} n + \frac{\ln B}{\theta_2 - \theta_1}$$

这个检验法可用图 18 表示如下.

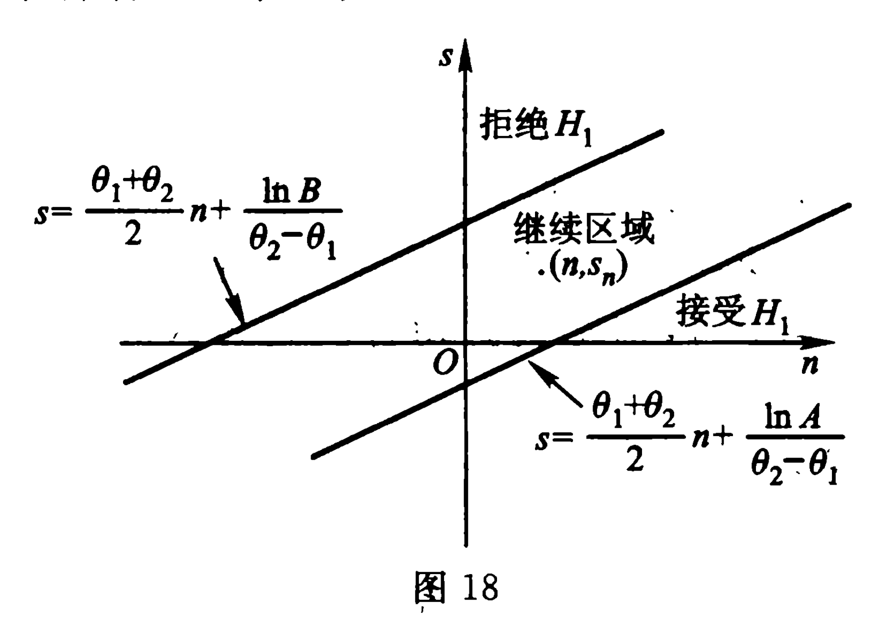


图 18

从上述例子看出，序贯概率比检验的实施并不复杂，它仅仅依赖于事先选定的常数 A 与 B. 怎样选定 A、B 呢? 这就需要仔细研究 SPRT 的性质. 很自然地有下列四个问题:

(1) 是否有限步后一定会停止观测(抽样)? 即是否成立 $P_i(\tau^* < \infty) = 1 (i=1,2)$，这里 $P_i(A)$表示 X 的真实分布密度(或概率函数)是 $f_i(x)$时事件 A 的概率.

(2) 如何计算 $P_i(\tau^* < \infty,\ \lambda_{\tau^*} \leqslant A)$. 通常记 $\alpha = P_1(\tau^* < \infty,\ \lambda_{\tau^*} \geqslant B)$, $\beta = P_2(\tau^* < \infty,\ \lambda_{\tau^*} \leqslant A)$. 它们是两类错误的概率. 我们希望找出 α、β 与 A、B 之间的关系，从而根据给定的 α、β 来确定 A、B.

(3) 如何计算平均样本量 $E_i\tau^* (i=1,2)$，这里 E_i 是与 P_i 相

应的数学期望. 自然希望 $E_i\tau^*$ 越小越好.

(4) SPRT $S(A,B)$有何优良性? 有无检验法比 SPRT 还要"优"?

怎样研究这些问题呢? 一个基本观点是:SPRT 的研究可化为随机游动的研究. 从(2.1)知

$$\ln \lambda_n = \sum_1^n \ln \frac{f_2(x_i)}{f_1(x_i)} \quad (n \geqslant 1)$$

令

$$z_i = \ln \frac{f_2(x_i)}{f_1(x_i)} \quad (n \geqslant 1)$$

$$Z_n = \ln \lambda_n = \sum_1^n z_i \quad (n \geqslant 1)$$

注意 $z_1, z_2, \cdots$是相互独立同分布随机变量列,$\{Z_n, n \geqslant 1\}$就是通常所谓的随机游动,而 $S(A,B)$的停止法则恰好是 $\tau^* = \inf\{n: n \geqslant 1, Z_n \overline{\in} (\ln A, \ln B)\}$,即随机游动首次离开区间$(\ln A, \ln B)$的时刻.

下面先对随机游动进行一番研究,然后再论述 SPRT 的性质.

引理 2.1(Wald 引理) 设 τ 是随机游动$\{Z_n, n \geqslant 1\}$的任一停止时间,即对一切 $n \geqslant 1$,有 Borel 集 B_n 使得

$$\{\tau \leqslant n\} = \{(Z_1, \cdots, Z_n) \in B_n\}$$

若 EZ_1 存在且 $E\tau < \infty$,则有

$$EZ_\tau = EZ_1 \cdot E\tau \tag{2.4}$$

证明:记 $x^+ = \max(x, 0)$, $x^- = \max(-x, 0)$,记

$$z_i = Z_i - Z_{i-1} (i \geqslant 1), \quad Z_0 \overset{\mathrm{d}}{=} 0$$

由于

$$\{\tau \geqslant i\} = \{(Z_1, \cdots, Z_{i-1}) \in B_{i-1}^0\}$$

故 $I(\tau \geqslant i)$与 z_i 相互独立(这里 I_A 表示 A 的示性函数),所以

$$E\left(\sum_{i=1}^{\tau} z_i^+\right) = E\left(\sum_{i=1}^{\infty} I_{\{\tau \geqslant i\}} \cdot z_i^+\right)$$

$$=\sum_{i=1}^{\infty}E(I_{(\tau\geqslant i)}\cdot z_i^+)$$

$$=\sum_{i=1}^{\infty}P(\tau\geqslant i)\cdot Ez_i^+=Ez_1^+\cdot E\tau$$

同理 $$E\left(\sum_{i=1}^{\tau}z_i^-\right)=Ez_1^-\cdot E\tau$$

由于 Ez_1^+、Ez_1^- 中至少一个有限，故

$$EZ_\tau=E\left(\sum_{i=1}^{\tau}(z_i^+-z_i^-)\right)=E\left(\sum_1^{\tau}z_i^+\right)-E\left(\sum_1^{\tau}z_i^-\right)$$

$$=EZ_1\cdot E\tau.$$ 证毕.

引理 2.2 设 τ 是随机游动 $\{Z_n,n\geqslant 1\}$ 的任一停止时间. 若 $EZ_1=0$，$EZ_1^2<\infty$，$E\tau<\infty$，则

$$EZ_\tau^2=EZ_1^2\cdot E\tau \tag{2.5}$$

证明：使用记号 $a\wedge b=\min(a,b)$. 易知

$$Z_{\tau\wedge n}=\sum_{j=1}^{n}I_{(\tau\geqslant j)}z_j\quad (z_j\overset{\mathrm{d}}{=}Z_j-Z_{j-1}，Z_0\overset{\mathrm{d}}{=}0)$$

任意固定 $n>m\geqslant 0$，易知

$$E(Z_{\tau\wedge n}-Z_{\tau\wedge m})^2=E\left(\sum_{j=m+1}^{n}I_{(\tau\geqslant j)}z_j\right)^2$$

$$=\sum_{j=m+1}^{n}E(I_{(\tau\geqslant j)}z_j^2)+2\sum_{m<i<j\leqslant n}E(I_{(\tau\geqslant j)}z_iz_j)$$

由于 $I_{(\tau\geqslant j)}$ 与 z_j 是相互独立的，故

$$E(I_{(\tau\geqslant j)}z_j^2)=Ez_1^2\cdot E(I_{(\tau\geqslant j)})$$

当 $j>i$ 时，

$$E(I_{(\tau\geqslant j)}z_iz_j)=E(I_{(\tau\geqslant j)}z_i)\cdot Ez_j=0$$

于是

$$E(Z_{\tau\wedge n}-Z_{\tau\wedge m})^2=Ez_1^2\cdot\sum_{j=m+1}^{n}E(I_{(\tau\geqslant j)})$$

$$=Ez_i^2(E(\tau\wedge n)-E(\tau\wedge m)) \tag{2.6}$$

令 $n\to\infty$ 得到

$$E(Z_\tau - Z_{\tau\wedge m})^2 \leqslant Ez_1^2(E\tau - E(\tau\wedge m))$$

于是

$$\lim_m E(Z_\tau - Z_{\tau\wedge m})^2 = 0.$$

从而

$$EZ_\tau^2 = \lim_m E(Z_{\tau\wedge m})^2$$

另一方面，在(2.6)中令 $m=0$ 得 $EZ_{\tau\wedge n}^2 = Ez_1^2 \cdot E(\tau\wedge n)$.

于是

$$EZ_\tau^2 = \lim_m Ez_1^2 \cdot E(\tau\wedge m) = EZ_1^2 \cdot E\tau$$

证毕.

给定区间(a,b)，$-\infty<a<0<b<\infty$. 我们来研究随机游动$\{Z_n, n\geqslant 1\}$首次离开区间(a,b)的时刻τ^*，即

$$\tau^* = \inf\{n: n\geqslant 1,\quad Z_n \overline{\in} (a,b)\} \tag{2.7}$$

引理 2.3(Stein)　设 $P(Z_1=0)<1$，则存在常数 $M>0$，$r>0$，使得

$$P(\tau^* > n) \leqslant Me^{-rn} \quad (\text{一切 } n) \tag{2.8}$$

证明：首先指出，存在 $m\geqslant 1$ 使得 $P(|Z_m|>b-a)\overset{d}{=}p>0$. 实际上，既然 $P(Z_1=0)<1$，故有 $\delta>0$ 满足 $P(|Z_1|>\delta)>0$，取 $m\geqslant\dfrac{b-a}{\delta}$就行了. 这是因为

$$\begin{aligned}P(Z_m > b-a) &\geqslant P(z_1>\delta, z_2>\delta,\cdots,z_m>\delta)\\ &= [P(z_1>\delta)]^m\end{aligned}$$

同理 $P(Z_m<-(b-a))\geqslant[P(z_1<-\delta)]^m$，故 $P(|Z_m|>b-a)>0$.

任给定 $n>m$，令 $q=\left[\dfrac{n}{m}\right]$，$I_0=(a,b)$，则

$$\begin{aligned}P(\tau^*>n) &= P(Z_1\in I_0,\cdots,Z_n\in I_0)\\ &\leqslant P(Z_m\in I_0,\ Z_{2m}\in I_0,\cdots,Z_{qm}\in I_0)\\ &\leqslant P(|Z_m|\leqslant b-a,\ |Z_{2m}-Z_m|\leqslant b-a,\cdots,\\ &\qquad |Z_{qm}-Z_{(q-1)m}|\leqslant b-a)\end{aligned}$$

$$=[P(|Z_m|\leqslant b-a)]^q$$

$$=(1-p)^q\leqslant(1-p)^{\frac{n}{m}-1}.$$

若 $p<1$,则

$$P(\tau^*>n)\leqslant\frac{1}{1-p}\exp\left\{-\frac{n}{m}\ln\frac{1}{1-p}\right\}\quad(n>m)$$

若 $p=1$,则

$$P(\tau^*>n)=0\quad(n>m)$$

总之存在 $M>0$, $r>0$ 使得(2.8)成立. 证毕.

引理 2.4 设 $P(Z_1=0)<1$,则 $P(\tau^*<\infty)=1$,且存在 $\lambda>0$ 使得 $Ee^{\lambda\tau^*}<\infty$,从而 τ^* 的一切阶矩均有限.

证明:可从(2.8)直接推出.

引理 2.5(序贯分析基本恒等式) 若 $\varphi(t)=Ee^{tZ_1}<\infty$, $P(Z_1=0)<1$ 且$\varphi(t)\geqslant1$(对某个 t),则

$$E(e^{tZ_{\tau^*}}/(\varphi(t))^{\tau^*})=1\tag{2.9}$$

其中 τ^* 由(2.7)确定.

证明:记 $z_k=Z_k-Z_{k-1}$, $Z_0=0$,则

$$\begin{aligned}E\left\{I_{(\tau^*\leqslant n)}\frac{e^{tZ_{\tau^*}}}{(\varphi(t))^{\tau^*}}\right\}&=\sum_{k=1}^{n}E\left\{I_{(\tau^*=k)}\frac{e^{tZ_k}}{(\varphi(t))^k}\right\}\\&=\sum_{k=1}^{n}\left\{E\left[I_{(\tau^*\geqslant k)}\frac{e^{tZ_k}}{(\varphi(t))^k}\right]-\right.\\&\qquad\left.E\left[I_{(\tau^*\geqslant k+1)}\frac{e^{tZ_k}}{(\varphi(t))^k}\right]\right\}\end{aligned}\tag{2.10}$$

由 $\{\tau^*\geqslant k\}=\{Z_1\in(a,b),\cdots,Z_{k-1}\in(a,b)\}$知随机变量

$$I_{(\tau^*\geqslant k)}\cdot\frac{e^{tZ_{k-1}}}{(\varphi(t))^{k-1}}\text{ 与 }\frac{e^{tz_k}}{\varphi(t)}$$

相互独立,从而

$$\begin{aligned}E\left[I_{(\tau^*\geqslant k)}\cdot\frac{e^{tZ_k}}{(\varphi(t))^k}\right]&=E\left[I_{(\tau^*\geqslant k)}\cdot\frac{e^{tZ_{k-1}}}{(\varphi(t))^{k-1}}\right]\cdot E\left[\frac{e^{tz_k}}{\varphi(t)}\right]\\&=E\left[I_{(\tau^*\geqslant k)}\cdot\frac{e^{tZ_{k-1}}}{(\varphi(t))^{k-1}}\right]\end{aligned}$$

再利用(2.10)我们得：

$$E\left[I_{(\tau^*\leqslant n)}\cdot\frac{e^{tZ_{\tau^*}}}{(\varphi(t))^{\tau^*}}\right]=1-E\left[I_{(\tau^*>n)}\cdot\frac{e^{tZ_n}}{(\varphi(t))^n}\right]$$

另一方面，当 $\tau^*>n$ 时，$|Z_n|\leqslant b-a$，又 $\varphi(t)\geqslant 1$，故

$$E\left[I_{(\tau^*>n)}\cdot\frac{e^{tZ_n}}{(\varphi(t))^n}\right]\leqslant e^{|t|(b-a)}\cdot P(\tau^*>n)\to 0\ (n\to\infty)$$

所以(2.9)成立. 证毕.

对于随机游动 $\{Z_n,n\geqslant 1\}$，我们最关心的是计算 $L=P(Z_{\tau^*}\leqslant a)$ 及 $E\tau^*$，这里 τ^* 是由(2.7)定义的. 要给出一般情形下的精确计算公式是困难的，但我们可给出有实用价值的近似公式.

设 $P(Z_1=0)<1$，有 $t_0\neq 0$ 使得 $\varphi(t_0)=Ee^{t_0Z_1}=1$，则有下列近似公式：

(1) $L\doteq\dfrac{e^{t_0b}-1}{e^{t_0b}-e^{t_0a}}$ (2.11)

这里 $L=P(Z_{\tau^*}\leqslant a)$；

(2) $EZ_{\tau^*}\doteq b+L(a-b)$； (2.12)

(3) 若 $EZ_1\neq 0$，那么

$$E\tau^*\doteq\frac{b+L(a-b)}{EZ_1}\tag{2.13}$$

(4) 若 $EZ_1=0$，$EZ_1^2<\infty$，那么

$$E\tau^*\doteq\frac{La^2+(1-L)b^2}{EZ_1^2}\tag{2.14}$$

这些近似公式是怎样推导出来的呢？由基本恒等式(2.9)知 $E(e^{t_0Z_{\tau^*}})=1$，在集合 $\{Z_{\tau^*}\leqslant a\}$ 上用 a 作为 Z_{τ^*} 的近似值，在集合 $\{Z_{\tau^*}\geqslant b\}$ 上用 b 作为 Z_{τ^*} 的近似值，于是有

$$\int_{\{Z_{\tau^*}\leqslant a\}}e^{t_0a}dP+\int_{\{Z_{\tau^*}\geqslant b\}}e^{t_0b}dP\doteq 1$$

由此即得到(2.11).

同理可得到(2.12). 再利用引理 2.1 和引理 2.2 即可得到(2.13)、(2.14).

现在利用随机游动的性质导出 SPRT 的统计特性，仍沿用以前的记号.

定理 2.1　对于 $i=1,2$ 均有

$$P_i(\tau^* < \infty)=1$$

$$E_i e^{\lambda\tau^*} < \infty \quad (\text{对某个 } \lambda > 0)$$

$$E_i(\tau^*)^k < \infty \quad (\text{一切 } k \geqslant 1)$$

证明：令 $z_i=\ln[f_2(x_i)/f_1(x_i)]$，$Z_n=\ln\lambda_n=z_1+\cdots+z_n$，$a=\ln A$，$b=\ln B$，则 $\{Z_n, n\geqslant 1\}$ 是随机游动且

$$\tau^*=\inf\{n: n\geqslant 1, Z_n \overline{\in} (a,b)\}$$

当 Z_1 是连续型随机变量时，$E_1 Z_1=\int f_1(x)\ln\frac{f_2(x)}{f_1(x)}\mathrm{d}x$，利用不等式 $\ln x<x-1(x>0$ 且 $x\neq 1)$，知

$$E_1 Z_1 < \int f_1(x)\left[\frac{f_2(x)}{f_1(x)}-1\right]\mathrm{d}x=0$$

故 $P_1(Z_1=0)<1$. 当 Z_1 是离散型时也有同样的结论. 利用引理 2.4 知

$$P_1(\tau^* < \infty)=1 \text{ 且 } E_1 e^{\lambda\tau^*} < \infty \quad (\text{某个 } \lambda > 0)$$

$$E_1(\tau^*)^k < \infty$$

同理知

$$P_2(\tau^* < \infty)=1 \text{ 且 } E_2 e^{\lambda\tau^*} < \infty \quad (\text{某个 } \lambda > 0)$$

$$E_2(\tau^*)^k < \infty$$

证毕.

定理 2.2　对于 $S(A,B)$ 而言，有下列关系式

$$\alpha \leqslant \frac{1}{B}(1-\beta), \quad \beta \leqslant A(1-\alpha) \tag{2.15}$$

证明：设 $f_1(x)$，$f_2(x)$ 是分布密度(当它们是概率函数时，下面的证明方法仍有效，但要将积分改为求和)，易知有集合 E_n (Borel 集)使得 $\{\tau^*=n\}=\{(x_1,\cdots,x_n)\in E_n\}$. 于是

$$\alpha=P_1(\lambda_{\tau^*}\geqslant B)=\sum_{n=1}^{\infty}P_1(\tau^*=n, \lambda_n \geqslant B)$$

$$=\sum_{n=1}^{\infty}\int_{E_n\cap\{\lambda_n\geqslant B\}}\prod_{1}^{n}f_1(x_i)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$=\sum_{n=1}^{\infty}\int_{E_n\cap\{\lambda_n\geqslant B\}}\frac{1}{\lambda_n}\prod_{1}^{n}f_2(x_i)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$\leqslant\frac{1}{B}\sum_{n=1}^{\infty}\int_{E_n\cap\{\lambda_n\geqslant B\}}\prod_{1}^{n}f_2(x_i)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$=\frac{1}{B}\sum_{n=1}^{\infty}P_2(\tau^*=n,\ \lambda_n\geqslant B)$$

$$=\frac{1}{B}P_2(\lambda_{\tau^*}\geqslant B)=\frac{1}{B}(1-\beta)$$

同理可证 $\beta\leqslant A(1-\alpha)$. 证毕.

从(2.15)知$\frac{\beta}{1-\alpha}\leqslant A<B\leqslant\frac{1-\beta}{\alpha}$,在实用上常用近似公式:

$$A\doteq\frac{\beta}{1-\alpha},\quad B\doteq\frac{1-\beta}{\alpha}\tag{2.16}$$

公式(2.16)的重要意义在于:对给定的 α、β,很容易确定 A、B 的值,从而得到所要的 SPRT. 在确定 A、B 时根本不需要了解总体的分布类型是什么,也不像普通的固定样本量检验法那样要查临界值表.

对于给定的 α、β,使用(2.16)可得 A、B 的近似值 $A^*=\frac{\beta}{1-\alpha}$, $B^*=\frac{1-\beta}{\alpha}$. 当然方案 $S(A^*,B^*)$的两类错误概率与 α、β 可能不同,但却有下列重要不等式.

定理 2.3 给定 $\alpha>0$,$\beta>0$(满足 $\alpha+\beta<1$),设

$$A^*=\frac{\beta}{1-\alpha},\quad B^*=\frac{1-\beta}{\alpha}$$

$S(A^*,B^*)$的两类错误概率分别是 α^*、β^*,则恒有

$$\alpha^*\leqslant\frac{\alpha}{1-\beta},\quad \beta^*\leqslant\frac{\beta}{1-\alpha}$$

$$\alpha^* + \beta^* \leqslant \alpha + \beta$$

证明:从条件 $\alpha+\beta<1$ 知 $A^*<B^*$,从定理 2.2 知

$$\alpha^* \leqslant \frac{1}{B^*}(1-\beta^*) = \frac{\alpha}{1-\beta}(1-\beta^*) \leqslant \frac{\alpha}{1-\beta}$$

$$\beta^* \leqslant A^*(1-\alpha^*) = \frac{\beta}{1-\alpha}(1-\alpha^*) \leqslant \frac{\beta}{1-\alpha}$$

于是

$$(1-\beta)\alpha^* \leqslant \alpha(1-\beta^*),\ (1-\alpha)\beta^* \leqslant \beta(1-\alpha^*)$$

将这两个不等式的两边分别相加,得 $\alpha^*+\beta^*\leqslant\alpha+\beta$. 证毕.

这个定理告诉我们,采用近似公式(2.16)后,两类错误的概率之和是不增大的.

设 $Z_1=\ln[f_2(x_1)/f_1(x_1)]$,若有 $h_i\neq 0$ 满足 $E_i e^{h_i Z_1}=1$,则从(2.11)得到

$$L_i \stackrel{\mathrm{d}}{=} P_i(\lambda_{\tau^*} \leqslant A) \doteq \frac{B^{h_i}-1}{B^{h_i}-A^{h_i}} \quad (i=1,2)$$

从而 $\alpha=1-L_1$ 和 $\beta=L_2$ 都可近似求出.

若取 $A\doteq\dfrac{\beta}{1-\alpha}$, $B\doteq\dfrac{1-\beta}{\alpha}$,从(2.13)得到

$$E_1\tau^* \doteq \frac{(1-\alpha)\ln\dfrac{\beta}{1-\alpha}+\alpha\ln\dfrac{1-\beta}{\alpha}}{E_1Z_1}$$

$$E_2\tau^* \doteq \frac{\beta\ln\dfrac{\beta}{1-\alpha}+(1-\beta)\ln\dfrac{1-\beta}{\alpha}}{E_2Z_1}$$

现在问:SPRT $S(A,B)$有何优良性? Wald 和 Wolfowicz(1948)证明了下述最优性:设 $S(A,B)$的两类错误概率是 α、β,样本量是 τ^*,则对任何序贯检验法 $\Delta=(\tau,d)$,只要 Δ 的两类错误概率 α^*、β^* 满足 $\alpha^*\leqslant\alpha$,$\beta^*\leqslant\beta$,就一定有 $E_i\tau\geqslant E_i\tau^*$ $(i=1,2)$,换句话说,$S(A,B)$的平均样本量达到最小.

这是现代统计学里最深刻的定理之一,严格证明相当难,读者可在[8]中找到它的完全证明.

以上讨论假设检验时,假定总体的真正分布密度(或概率函数)是 $f_1(x)$ 或 $f_2(x)$,在实际问题里,很可能真正的分布密度(或概率函数)既不是 $f_1(x)$,也不是 $f_2(x)$,而是 $f_0(x)$,此时如何计算 $S(A,B)$ 的接受假设 H_1 的概率及平均样本量呢?这个问题可以提成下列一般形式:设 $x_1,x_2,\cdots$ 是独立同分布的随机变量列,x_1 的分布密度(或概率函数,下同)是 $f(x,\theta)$,其中 $\theta\in\Theta$,Θ 是已知集合,至少含有两个元素.我们假定不同的 θ 对应不同的概率分布.研究检验问题

$$H_1:\theta=\theta_1\leftrightarrow H_2:\theta=\theta_2 \tag{2.17}$$

(H_1 是待检验的假设,H_2 是备择假设).我们使用 $S(A,B)$,其停止法则是

$$\tau^*=\inf\{n:n\geqslant 1,\quad \lambda_n\overline{\in}(A,B)\}$$

这里

$$\lambda_n=\prod_1^n f(x_i,\theta_2)\Big/\prod_1^n f(x_i,\theta_1),\quad n\geqslant 1$$

给定 $\theta\in\Theta$, $L(\theta)\overset{\mathrm{d}}{=}P_\theta$(接受 H_1)(这里 P_θ 乃是与 $f(x,\theta)$ 相应的概率),叫做检验的操作特性函数(OC 函数).$E_\theta\tau^*$ 是真正密度为 $f(x,\theta)$ 时的平均样本量(ASN 函数).怎样计算 $L(\theta)$ 与 $E_\theta\tau^*$ 呢?

令

$$z_i=\ln[f(x_i,\theta_2)/f(x_i,\theta_1)],\quad Z_n=\sum_1^n z_i(n\geqslant 1)$$

则 $\{Z_n,\ n\geqslant 1\}$ 是 P_θ 下的随机游动.设有 $h(\theta)\neq 0$ 满足 $E_\theta\mathrm{e}^{h(\theta)Z_1}=1$,则从(2.10)~(2.13)得到下列近似公式:

(1) $L(\theta)\doteq\dfrac{B^{h(\theta)}}{B^{h(\theta)}-A^{h(\theta)}}$

(2) 若 $E_\theta Z_1\neq 0$,那么

$$E_\theta\tau^*\doteq\frac{L(\theta)\ln A+(1-L(\theta))\ln B}{E_\theta Z_1}$$

(3) 若 $E_\theta Z_1=0$; $E_\theta Z_1^2<\infty$,那么

$$E_\theta \tau^* \doteq \frac{L(\theta)(\ln A)^2+(1-L(\theta))(\ln B)^2}{E_\theta Z_1^2}$$

下面对两个重要分布类型给出具体计算公式.

(一) 伯努利分布情形

设 $x_1,x_2,\cdots$是独立同分布随机变量列,x_i 取值 0 或 1,概率函数是

$$f(x,p)=p^x(1-p)^{1-x} \quad (x=0,1,0<p<1)$$

设假设 H_i 是 $p=p_i(i=1,2,p_1<p_2)$. 对于检验问题:

$$H_1 \leftrightarrow H_2$$

我们采用检验法 $S(A,B)$,它的停止法则是 τ^*. 我们来计算 OC 函数 $L(p)=P_p$(接受 H_1)与 ASN 函数 $E_p\tau^*$. 这里 P_p 是与 $f(x,p)$相应的概率,E_p 是相应的数学期望.

由于

$$Z_1=x_1\ln\frac{p_2}{p_1}+(1-x_1)\ln\frac{1-p_2}{1-p_1}$$

$$E_pZ_1=p\ln\frac{p_2}{p_1}+(1-p)\ln\frac{1-p_2}{1-p_1}$$

$$E_p\mathrm{e}^{hZ_1}=(1-p)\left(\frac{1-p_2}{1-p_1}\right)^h+p\left(\frac{p_2}{p_1}\right)^h$$

可以证明存在 $h=h(p)\neq 0$ 满足

$$E_p\mathrm{e}^{hZ_1}=1$$

于是

$$L(p)\doteq\frac{B^{h(p)}-1}{B^{h(p)}-A^{h(p)}}$$

若

$$E_pZ_1=p\ln\frac{p_2}{p_1}+(1-p)\ln\frac{1-p_2}{1-p_1}\neq 0$$

则

$$E_p\tau^* \doteq \frac{L(p)\ln A + (1-L(p))\ln B}{p\ln\dfrac{p_2}{p_1} + (1-p)\ln\dfrac{1-p_2}{1-p_1}}$$

若 $E_pZ_1=0$,则

$$E_p\tau^* \doteq \frac{L(p)(\ln A)^2 + (1-L(p))(\ln B)^2}{p\left(\ln\dfrac{p_2}{p_1}\right)^2 + (1-p)\left(\ln\dfrac{1-p_2}{1-p_1}\right)^2}$$

下面是数值例子,若 $p_1=0.05$, $p_2=0.10$, $\alpha=\beta=0.05$,对于检验问题 $p=p_1\leftrightarrow p=p_2$,取 $A=\dfrac{\beta}{1-\alpha}$, $B=\dfrac{1-\beta}{\alpha}$,则 $A=\dfrac{1}{19}$,$B=19$ 且

$$\tau^*=\inf\{n:n\geqslant 1,\ s_n\overline{\in}(0.72n-3.94,\ 0.72n+3.94)\}$$

这里

$$s_n=\sum_1^n x_i$$

$$L(p)\doteq\frac{19^h-1}{19^h-19^{-h}}$$

其中 $h=h(p)$满足方程

$$p=\frac{1-\left(\dfrac{18}{19}\right)^h}{2-\left(\dfrac{18}{19}\right)^h}$$

可算出 $E_{0.05}\tau^*\doteq159$, $E_{0.10}\tau^*\doteq128$.

由此看出,SPRT 的平均样本量是相当大的,然而可以验证它比固定样本量的检验法要好.

仍令 $\alpha=\beta=0.05$,我们来计算一下,为了达到同样的 α、β,固定样本量方法的样本量要多大才行. 这时样本量肯定很大,可用极限定理来估计. 在 $p=p_1=0.05$ 的条件下,$s_n=\sum\limits_1^n x_i$ 近似服从 $N(0.05n,\ 0.0475n)$,在 $p=p_2=0.10$ 的条件下,s_n 近似服从 $N(0.10n,\ 0.09n)$,找 λ 和 n 满足

$$P_{0.05}\left(\frac{1}{n}s_n \geqslant \lambda\right) \doteq 0.05,\ P_{0.10}\left(\frac{1}{n}s_n < \lambda\right) \doteq 0.05$$

显然 λ 和 n 应由下式确定：

$$\frac{(\lambda - 0.05)\sqrt{n}}{\sqrt{0.0475}} = 1.65$$

$$\frac{(\lambda - 0.10)\sqrt{n}}{\sqrt{0.09}} = -1.65$$

由此知 $n \doteq 292$，$\lambda \doteq 0.072$，$\lambda \cdot n = 21$，故 $s_{292} \geqslant 21$ 时拒绝 H_1，$s_{292} < 21$ 时接受 H_1，这时两类错误概率都是 0.05，但样本量却是 SPRT 的平均样本量 $\left(\frac{1}{2}(E_{0.05}\tau^* + E_{0.10}\tau^*) = 144\right)$ 的二倍.

（二）正态分布情形

设 $x_1, x_2, \cdots$ 是独立同分布随机变量列，$x_1 \sim N(\theta, 1)$. 假设 H_i 是"$\theta = \theta_i$"($i = 1, 2, \theta_1 < \theta_2$)，对于检验问题：

$$H_1 : \theta = \theta_1 \leftrightarrow H_2 : \theta = \theta_2$$

用检验法 $S(A, B)$. 我们计算 $L(\theta)$ 及 $E_\theta \tau^*$.

$$Z_1 \overset{\mathrm{d}}{=} \ln[f(x_1, \theta_2)/f(x_1, \theta_1)]，其中$$

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \theta)^2\right\}$$

易知

$$Z_1 = (\theta_2 - \theta_1)x_1 + \frac{1}{2}(\theta_1^2 - \theta_2^2)$$

故

$$E_\theta e^{hZ_1} = \exp\left\{\frac{h}{2}(\theta_2 - \theta_1) \cdot [2\theta + h(\theta_2 - \theta_1) - (\theta_1 + \theta_2)]\right\}$$

可见 $h = h(\theta) = \dfrac{\theta_1 + \theta_2 - 2\theta}{\theta_2 - \theta_1}$ 时 $E_\theta e^{hZ_1} = 1$，于是 $\theta \neq \dfrac{\theta_1 + \theta_2}{2}$ 时

$$L(\theta) = \frac{B^{h(\theta)} - 1}{B^{h(\theta)} - A^{h(\theta)}}$$

当 $\theta = \theta_0 = \dfrac{\theta_1 + \theta_2}{2}$ 时，利用连续性知

$$L(\theta_0) \doteq \lim_{\theta \to \theta_0} \frac{B^{h(\theta)} - 1}{B^{h(\theta)} - A^{h(\theta)}}$$

$$= \frac{\ln B}{\ln B - \ln A}$$

当 $\theta \neq \frac{\theta_1 + \theta_2}{2}$ 时，$E_\theta Z_1 \neq 0$，故

$$E_\theta \tau^* \doteq \frac{L(\theta)\ln A + (1 - L(\theta))\ln B}{(\theta_2 - \theta_1)\theta + \frac{1}{2}(\theta_1^2 - \theta_2^2)}$$

当 $\theta = \frac{\theta_1 + \theta_2}{2}$ 时，$E_\theta Z_1 = 0$，$E_\theta Z_1^2 = (\theta_2 - \theta_1)^2$，故

$$E_\theta \tau^* \doteq \frac{-(\ln A)(\ln B)}{(\theta_2 - \theta_1)^2}$$

取 $A = \frac{\beta}{1-\alpha}$，$B = \frac{1-\beta}{\alpha}$，易知

$$E_{\theta_2} \tau^* \doteq \frac{\beta \ln \frac{\beta}{1-\alpha} + (1-\beta)\ln \frac{1-\beta}{\alpha}}{\frac{1}{2}(\theta_2 - \theta_1)^2} \tag{2.18}$$

$$E_{\theta_1} \tau^* \doteq \frac{(1-\alpha)\ln \frac{\beta}{1-\alpha} + \alpha \ln \frac{1-\beta}{\alpha}}{-\frac{1}{2}(\theta_2 - \theta_1)^2} \tag{2.19}$$

下面我们比较 SPRT 和固定样本量检验法，以显示前者在节约样本量方面的好处.

根据 N－P 引理，最优的固定样本量检验是这样的：当 $\bar{x} = \frac{1}{n}\sum_1^n x_i > C$ 时拒绝 $H_1: \theta = \theta_1$，当 $\bar{x} \leqslant C$ 时接受 H_1. 这里的 n，C 应使得

$$P_{\theta_1}(\bar{x} > C) = \alpha, \quad P_{\theta_2}(\bar{x} \leqslant C) = \beta$$

设 K_λ 满足 $\int_{K_\lambda}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \lambda$，则

$$(C - \theta_1)n = K_\alpha, (C - \theta_2)n = K_{1-\beta}$$

于是

$$n \doteq \frac{(K_\alpha - K_{1-\beta})^2}{(\theta_2 - \theta_1)^2}$$

$$C = \theta_1 + \frac{(\theta_2 - \theta_1) K_\alpha}{|K_\alpha - K_{1-\beta}|}$$

取 $A = \frac{\beta}{1-\alpha}$, $B = \frac{1-\beta}{\alpha}$, $S(A,B)$的两类错误的概率近似地等于α、β,令

$$\mu_i = \frac{E_{\theta_i}\tau^*}{n} \quad (i=1,2)$$

从(2.18)、(2.19)得

$$\mu_1 \doteq -2\frac{\alpha \ln \frac{1-\beta}{\alpha} + (1-\alpha)\ln \frac{\beta}{1-\alpha}}{(K_\alpha - K_{1-\beta})^2}$$

$$\mu_2 \doteq 2\frac{(1-\beta)\ln \frac{1-\beta}{\alpha} + \beta \ln \frac{\beta}{1-\alpha}}{(K_\alpha - K_{1-\beta})^2}$$

对于常用的几个α、β,表 2.1,表 2.2 列出了 $100\mu_1$ 和 $100\mu_2$ 的值.

表 2.1　$100\mu_1$

β	α			
	0.01	0.03	0.05	0.10
0.01	41.5	48.8	52.5	56.4
0.03	38.4	46.2	50.5	55.7
0.05	36.6	44.7	49.2	55.2
0.10	33.1	41.4	46.4	53.7

表 2.2　$100\mu_2$

β	α			
	0.01	0.03	0.05	0.10
0.01	41.5	38.4	36.6	33.1
0.03	48.8	46.2	44.7	41.4
0.05	52.5	50.5	49.2	46.4
0.10	56.4	55.7	46.2	53.7

从这两张表可看出，SPRT 所需要的平均样本量比固定样本量的最优检验的样本量要小得多，使用前者平均可节省样本量一半左右.

例 2.3 设 $x_1 \sim N(\theta,1)$，$H_1:\theta=0$，$H_2:\theta=1$，$\alpha=\beta=0.01$，取 $A=\dfrac{\beta}{1-\alpha}=\dfrac{1}{99}$，$B=\dfrac{1-\beta}{\alpha}=99$，这时，

$$Z_1=x_1-\frac{1}{2},\quad E_0Z_1=-\frac{1}{2},\quad E_1Z_1=\frac{1}{2}$$

$$E_0\tau^* \doteq \frac{0.99\ln(99)^{-1}+0.01\ln 99}{-\dfrac{1}{2}} \doteq 9$$

$$E_1\tau^* \doteq \frac{0.01\ln(99)^{-1}+0.99\ln 99}{\dfrac{1}{2}} \doteq 9$$

若采用固定样本量的最优检验法，则应取 $n\doteq(K_{0.01}-K_{0.99})^2$，查正态分布表知 $K_{0.01}=2.33$，$K_{0.99}=-2.33$，于是 $n\doteq 22$. 可见 SPRT 所需的平均样本量是非序贯情形最优检验法的样本量的 41%.

以上的讨论都是针对简单假设和简单备择假设进行的，这时的 SPRT 有很大的优点. 对于复杂假设，应采取什么样的序贯检验法呢？这个问题现代有大量研究，但未能彻底解决. 下面只对一种最简单的情况介绍处理方法.

设 X 的分布函数是 $F(x,\theta)$，分布密度（或概率函数）是 $f(x,\theta)$，这里 $\theta\in(\underline{\theta},\overline{\theta})=\Theta$，$-\infty\leqslant\underline{\theta}<\overline{\theta}\leqslant\infty$. 给定 $\theta_1<\theta_2$（$\theta_i\in\Theta$，$i=1,2$），考虑检验问题：

$$H_1:\theta\leqslant\theta_1 \leftrightarrow H_2:\theta\geqslant\theta_2$$

如何进行检验？设 $x_1,x_2,\cdots$ 是 X 的观测序列，仍令

$$\lambda_n=\prod_1^n f(x_i,\theta_2)/f(x_i,\theta_1)$$

取 $0<A<1<B$，仍采用 $S(A,B)$，即停止法则为

$$\tau^* = \inf\{n: n \geqslant 1, \lambda_n \overline{\in} (A, B)\}$$

判决法则为：

$$d^* = \begin{cases} 拒绝 H_1, & 当 \lambda_{\tau^*} \geqslant B \\ 接受 H_1, & 当 \lambda_{\tau^*} \leqslant A \end{cases}$$

这时检验法 $S(A, B)$有何统计特性呢？我们只考虑$\{F(x,\theta)\}$是随机增的情况.

定义 2.2 称分布族$\{F(x,\theta)\}$是随机增的，若对一切$\theta_1 < \theta_2$有

$$F(x,\theta_1) \geqslant F(x,\theta_2) \quad (一切\ x)$$

随机增的分布族是很不少的，例如指数分布族

$$F(x,\theta) = \delta(x)(1 - e^{-\frac{1}{\theta}x}), \quad \theta \in (0, \infty)$$

其中

$$\delta(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

和方差为 1 的正态分布族

$$F(x,\theta) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(u-\theta)^2} du, \quad \theta \in (-\infty, \infty)$$

都是随机增的.

可以证明下列重要结论：若$\{F(x,\theta)\}$是随机增的，则操作特性函数 $L(\theta) \stackrel{d}{=} P_\theta(接受\ H_1)$是$\theta$的不增函数. 因此有

$$\sup_{\theta \leqslant \theta_1} P_\theta(拒绝\ H_1) = P_{\theta_1}(拒绝\ H_1)$$

$$\sup_{\theta \geqslant \theta_2} P_\theta(接受\ H_1) = P_{\theta_2}(接受\ H_1)$$

至于平均样本量$E_\theta \tau^*$，作为θ的函数相当复杂.

作为本节的末尾，我们指出，SPRT 虽然在节省样本量上有很大优越性，但样本量并无固定的上界，这对实际工作带来不便. 30 年来，许多截尾型的序贯检验法（这时样本量有上界）受到广泛的重视，如著名的美国军用标准 781 和我国军用标准 899 就是采用截尾型的序贯检验法. 参看[22].

我们指出，上述 SPRT 在应用中还有一种局限性："序贯性太强"——样本是对个体一个一个地逐次观测得到的（一个个体观测完毕才可能开始观测下一个），不能同时进行观测．对于那些需要观测时间很长才能得到个体观测值的实际工作（例如许多医药疗效试验），SPRT 很难实施．针对这种情况，后来出现了多种新的序贯检验法：成组序贯检验（Group sequential test）、时间序贯检验（Time sequential test）等，读者可参阅[14].

§3 序贯估计与随机逼近

序贯估计就是用序贯方法估计总体的特性值（比如估计总体所含的未知参数的值）．与序贯检验相类似，它的重要性也是在两个方面．一方面，序贯估计在许多情况下比起非序贯估计来有较高的效率，即为了达到同样的估计精度，序贯方法所需的（平均）样本量要小些；另一方面，在某些情况下不存在非序贯的估计满足所提的要求，但有序贯估计合乎要求．这后一方面告诉我们，序贯估计是不可少的.

序贯估计包含序贯点估计与序贯区间估计．序贯估计的理论和方法从 20 世纪 50 年代以来有很大发展，研究的问题很多，方法也多种多样，尚缺乏统一的方法和理论．限于篇幅，我们不去对序贯估计作全面论述，只以正态分布均值的估计问题为例，说明某些重要的概念和方法.

设 $X\sim N(\mu,\sigma^2)$，μ、σ^2 都未知，X_1、X_2，…是 X 的独立观测序列，我们来研究 μ 的估计问题.

问题：给定 $l>0$ 及 $\gamma\in(0,1)$，是否存在正整数 n 及 $\varphi_1(X_1,\cdots,X_n)$，$\varphi_2(X_1,\cdots,X_n)$满足

$$0\leqslant\varphi_2-\varphi_1\leqslant l$$

$$P_{\mu\sigma}(\varphi_1\leqslant\mu\leqslant\varphi_2)\geqslant\gamma\quad(\text{一切 }\mu,\sigma)$$

换句话说，是否存在固定宽度的置信区间（这里 $P_{\mu\sigma}$ 乃是参数

为μ、σ时的概率)?

这个问题的答案是否定的,我们可以证明下列一般性结论.

定理 3.1 设 X 的分布密度是$\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$,其中 $\mu\in(-\infty,\infty)$,$\sigma\in(0,\infty)$都是未知参数,$X_1,\cdots,X_n$ 是 X 的样本. 则对任何 n 及 $l>0$,不存在 μ 之长度不超过 l 的置信区间(置信水平是 γ),即不存在 $\varphi_1(X_1,\cdots,X_n)\leqslant\varphi_2(X_1,\cdots,X_n)$使得

$$\varphi_2-\varphi_1\leqslant l \tag{3.1}$$

$$P_{\mu\sigma}(\mu\in[\varphi_1,\varphi_2])\geqslant\gamma\quad(\text{一切 }\mu,\sigma) \tag{3.2}$$

这里 $P_{\mu\sigma}$是分布密度为$\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$时相应的概率.

证明:设有 φ_1,φ_2 满足(3.1),我们指出必有 μ,σ 使得

$$P_{\mu\sigma}(\mu\in[\varphi_1,\varphi_2])<\gamma$$

实际上,令 $\delta=\frac{1}{2}(\varphi_1+\varphi_2)$,则

$$\{\mu\in[\varphi_1,\varphi_2]\}\subset\left\{|\delta-\mu|\leqslant\frac{1}{2}l\right\}$$

取正整数 $N>\frac{1}{\gamma}$,再取 $\mu_1<\mu_2<\cdots<\mu_{2N}$ 满足 $\mu_{i+1}-\mu_i>l$ $(i=1,2,\cdots,2N-1)$,于是集合

$$S_i=\left\{(x_1,\cdots,x_n):|\delta(x_1,\cdots,x_n)-\mu_i|\leqslant\frac{l}{2}\right\}$$

$(i=1,2,\cdots,2N)$两两不交. 因此

$$|P_{\mu_i\sigma}((X_1,\cdots,X_n)\in S_i)-P_{\mu_1\sigma}((X_1,\cdots,X_n)\in S_i)|$$

$$=\left|\int_{S_i}\cdots\int\frac{1}{\sigma^n}\prod_1^n f\left(\frac{x_k-\mu_i}{\sigma}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n-\right.$$

$$\left.\int_{S_i}\cdots\int\frac{1}{\sigma^n}\prod_1^n f\left(\frac{x_k-\mu_1}{\sigma}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n\right|$$



令

$$\frac{x_k-\mu_1}{\sigma}=y_k \quad (k=1,\cdots,n)$$

$$S_i^*=\{(y_1,\cdots,y_n);(x_1,\cdots,x_n)\in S_i\}$$

则

$$|P_{\mu_i\sigma}((X_1,\cdots,X_n)\in S_i)-P_{\mu_1\sigma}((X_1,\cdots,X_n)\in S_i)|$$

$$\leqslant\int_{S_i^*}\cdots\int\left|\prod_{k=1}^n f\left(y_k+\frac{\mu_1-\mu_i}{\sigma}\right)-\prod_1^n f(y_k)\right|\mathrm{d}y_1\cdots\mathrm{d}y_n$$

$$\leqslant\int_{R^n}\cdots\int\left|\prod_1^n f\left(y_k+\frac{\mu_1-\mu_i}{\sigma}\right)-\prod_1^n f(y_k)\right|\mathrm{d}y_1\cdots\mathrm{d}y_k$$

$$=\int_{R^n}\cdots\int|g(y+a)-g(y)|\mathrm{d}y$$

其中

$$y=(y_1,\cdots,y_n)$$

$$a=\frac{\mu_1-\mu_i}{\sigma}(1,1,\cdots,1)$$

$$g(y)=\prod_1^n f(y_k)$$

由于 $g\in L_1$，根据 L_1 连续性知 $a\to 0$ 时

$$\int_{R^n}\cdots\int|g(y+a)-g(y)|\mathrm{d}y\to 0$$

（见周民强编《实变函数》第 140 页，北京大学出版社，1985）. 于是

$$\lim_{\sigma\to\infty}|P_{\mu_i\sigma}((X_1,\cdots,X_n)\in S_i)-P_{\mu_1\sigma}((X_1,\cdots,X_n)\in S_i)|=0$$

$$(i=1,2,\cdots,2N)$$

可见存在 $\sigma_0>0$，对一切 $i=1,2,\cdots,2N$ 有

$$|P_{\mu_i\sigma_0}((X_1,\cdots,X_n)\in S_i)-P_{\mu_1\sigma_0}((X_1,\cdots,X_n)\in S_i)|\leqslant\frac{1}{2N}$$

但存在 i_0 使得

$$P_{\mu_1\sigma_0}((X_1,\cdots,X_n)\in S_{i_0})\leqslant\frac{1}{2N}$$

故

$$P_{\mu_{i_0}\sigma_0}((X_1,\cdots,X_n)\in S_{i_0})\leqslant\frac{1}{N}$$

即有

$$P_{\mu_{i_0}\sigma_0}\left\{|\delta-\mu_{i_0}|<\frac{l}{2}\right\}\leqslant\frac{1}{N}<\gamma$$

更有

$$P_{\mu_{i_0}\sigma_0}\{\mu_{i_0}\in[\varphi_1,\varphi_2]\}<\gamma$$

这表明(3.2)不成立. 证毕.

下面指出,对于正态分布 $N(\mu,\sigma^2)$,采用 C. Stein 提出的两阶段方法可得到 μ 的固定宽度的置信区间.

从定理 3.1 的证明过程看出,由于 σ 可以任意大,造成固定样本量方法出问题. 如果 σ 已知,则取

$$n\geqslant\frac{1}{d^2}\sigma^2u_\alpha^2\quad\left(\alpha=\frac{1+\gamma}{2}\right)$$

$$\varphi_1=\bar{x}_n-d,\ \varphi_2=\bar{x}_n+d\quad\left(d=\frac{l}{2}\right)$$

那么(3.1)和(3.2)就满足了,这里 u_α 为 $N(0,1)$ 的 α 分位点. 现在 σ 未知,Stein 的方法是先取一个样本来估计 σ 的值,然后再利用这个估计值确定第二个样本量的大小,具体叙述如下.

取 $m\geqslant2$,令 $\bar{x}_m=\frac{1}{m}\sum\limits_1^m x_i$,

$$\nu_m=\left[\frac{1}{m-1}\sum_{i=1}^m(x_i-\bar{x}_m)^2\right]^{\frac{1}{2}}$$

$$\tau_1=\inf\left\{n:n\geqslant m\text{ 且 }n\geqslant\frac{\nu_m^2}{d^2}\left[t_{\frac{1+\gamma}{2}}(m-1)\right]^2\right\}\tag{3.3}$$

其中 $d=\frac{l}{2}$, $t_\alpha(m-1)$是 $m-1$ 个自由度的 t 分布的 α 分位数.

显然 $P_{\mu\sigma}(\tau_1<\infty)=1$,区间

$$I=\left(\bar{x}_{\tau_1}-\frac{\nu_m}{\sqrt{\tau_1}}t_{\frac{1+\gamma}{2}}(m-1),\ \bar{x}_{\tau_1}+\frac{\nu_m}{\sqrt{\tau_1}}t_{\frac{1+\gamma}{2}}(m-1)\right)\tag{3.4}$$

的长度等于$\frac{2\nu_m}{\sqrt{\tau_1}}t_{\frac{1+\gamma}{2}}(m-1)\leqslant 2d=l$,我们将证明

$$P_{\mu\sigma}(\mu\in I)=\gamma \quad (对一切\ \mu,\sigma)$$

引理 3.1 对一切 $n\geqslant 2$, $\bar{x}_n$ 与$(\nu_2^2,\cdots,\nu_n^2)$是独立的,这里

$$\nu_i^2=\frac{1}{i-1}\sum_{k=1}^{i}(x_k-\bar{x}_i)^2 \quad (i\geqslant 2)$$

$$\bar{x}_i=\frac{1}{i}\sum_{k=1}^{i}x_k$$

证明:令

$$u_i=\frac{1}{\sigma}(x_i-\mu) \quad (i\geqslant 1)$$

则 $u_1,\cdots,u_n$ 独立同分布,$u_1\sim N(0,1)$. 令

$$y_i=\frac{1}{\sqrt{i(i+1)}}(u_1+\cdots+u_i-iu_{i+1}) \quad (i=1,2,\cdots,n-1)$$

$$y_n=\frac{1}{\sqrt{n}}(u_1+\cdots+u_n)$$

易知 $y_1,\cdots,y_n$ 独立同分布,$y_1\sim N(0,1)$,可以证明:对一切 $i\geqslant 2$,

$$\nu_i^2=\frac{\sigma^2}{i-1}(y_1^2+\cdots+y_{i-1}^2) \tag{3.5}$$

事实上,固定 $i\geqslant 2$,令 $\tilde{y}_i=\frac{1}{\sqrt{i}}(u_1+\cdots+u_i)$,则从 $u_1,\cdots,u_i$ 到 $y_1,\cdots,y_{i-1},\tilde{y}_i$ 的变换是正交变换,故

$$y_1^2+\cdots+y_{i-1}^2+\tilde{y}_i^2=\sum_1^i u_j^2$$

所以

$$y_1^2+\cdots+y_{i-1}^2=\sum_1^i u_j^2-i(\bar{u}_i)^2=\frac{1}{\sigma^2}\sum_{j=1}^{i}(x_j-\bar{x}_i)^2$$

从而(3.5)成立. 由于$(\nu_2^2,\cdots,\nu_n^2)$只与$y_1,\cdots,y_{n-1}$有关,而$\bar{x}_n=\frac{\sigma}{\sqrt{n}}y_n+\mu$只与$y_n$有关,故$\bar{x}_n$与$(\nu_2^2,\cdots,\nu_n^2)$相互独立. 证毕.

****引理 3.2** 设τ只取不小于m的整数值,且对一切$n\geqslant m$,存在B_n(Borel 集)满足

$$\{\tau=n\}=\{(\nu_2^2,\cdots,\nu_n^2)\in B_n\},$$

则有下列结论:

(1) $\bar{x}_\tau$是μ的无偏估计;

(2) $\frac{1}{\sigma}\sqrt{\tau}(\bar{x}_\tau-\mu)\sim N(0,1)$;

(3) $\frac{1}{\nu_m}\sqrt{\tau}(\bar{x}_\tau-\mu)\sim t(m-1)$.

证明:$E_{\mu\sigma}\bar{x}_\tau=\sum\limits_{n=m}^{\infty}E_{\mu\sigma}(I_{(\tau=n)}\bar{x}_n)$

$$=\sum_{n=m}^{\infty}P_{\mu\sigma}(\tau=n)\cdot E_{\mu\sigma}\bar{x}_n=\mu$$

可见$\bar{x}_\tau$是μ的无偏估计.

$$P_{\mu\sigma}\left(\frac{1}{\sigma}\sqrt{\tau}(\bar{x}_\tau-\mu)\leqslant x,\ \frac{1}{\sigma}\nu_m\leqslant y\right)$$

$$=\sum_{n=m}^{\infty}P_{\mu\sigma}\left(\tau=n,\ \frac{1}{\sigma}\nu_m\leqslant y,\ \frac{\sqrt{n}(\bar{x}_n-\mu)}{\sigma}\leqslant x\right)$$

$$=\sum_{n=m}^{\infty}P_{\mu\sigma}\left(\tau=n,\ \frac{1}{\sigma}\nu_m\leqslant y\right)P_{\mu\sigma}\left(\frac{\sqrt{n}(\bar{x}_n-\mu)}{\sigma}\leqslant x\right)$$

$$=\sum_{n=m}^{\infty}P_{\mu\sigma}\left(\tau=n,\ \frac{1}{\sigma}\nu_m\leqslant y\right)\Phi(x)$$

$$=P_{\mu\sigma}\left(\frac{1}{\sigma}\nu_m\leqslant y\right)\Phi(x)$$

其中$\Phi(x)$是$N(0,1)$的分布函数. 令$y\to\infty$得

$$P_{\mu\sigma}\left(\frac{1}{\sigma}\sqrt{\tau}(\bar{x}_\tau-\mu)\leqslant x\right)=\Phi(x)$$

故$\frac{1}{\sigma}\sqrt{\tau}(\bar{x}_\tau-\mu)$与$\frac{1}{\sigma}\nu_m$相互独立,从而

$$\frac{1}{\nu_m}\sqrt{\tau}(\bar{x}_\tau-\mu)\sim t(m-1)$$

证毕.

定理 3.2 设 τ_1 如(3.3)所定义,I 如(3.4)所定义,则

(1) $P_{\mu\sigma}(\mu\in I)=\gamma$;

(2) $E_{\mu\sigma}(\bar{x}_{\tau_1}-\mu)^2\leqslant\frac{(m-1)d^2}{(m-3)t^2}$ (当 $m>3$ 时)

其中 $d=\frac{l}{2}$, $t=t_{\frac{1+\gamma}{2}}(m-1)$.

证明:从引理 3.2 知

$$P_{\mu\sigma}\left(\left|\frac{1}{\nu_m}\sqrt{\tau_1}(\bar{x}_{\tau_1}-\mu)\right|\leqslant t_{\frac{1+\gamma}{2}}(m-1)\right)=\gamma$$

即有 $P_{\mu\sigma}(\mu\in I)=\gamma$.

$$E_{\mu\sigma}(\bar{x}_{\tau_1}-\mu)^2=E_{\mu\sigma}\left(\frac{\tau_1(\bar{x}_{\tau_1}-\mu)^2}{\nu_m^2}\cdot\frac{\nu_m^2}{\tau_1}\right)$$

$$\leqslant\frac{d^2}{t^2}E_{\mu\sigma}\left(\frac{\tau_1(x_{\tau_1}-\mu)^2}{\nu_m^2}\right)=\frac{d^2}{t^2}\frac{m-1}{m-3}$$

证毕.

定理 3.2 告诉我们,用 Stein 方法的确可以得到宽度不超过给定值的置信区间. 这个方法的缺点是样本量 τ_1 的大小不能控制,当 σ 很大时,$E_{\mu\sigma}\tau_1$ 可以取非常大的值. 第一阶段的样本量 m 应如何确定也没有回答. 如何改进 Stein 的方法一直是现代许多学者的研究对象.

例 3.1 已知总体 X 服从 $N(\mu,\sigma^2)$,μ,σ 未知. 设 $l=0.6$,$\gamma=0.95$. 我们可用 Stein 两阶段方法找出 μ 的置信水平为 0.95、长度不超过 0.6 的置信区间. 从总体中抽取样本量 $m=21$ 的样本(第一阶段样本):

0.72	1.69	0.82	0.21	−0.09	1.57	−0.26
1.08	1.94	1.60	1.19	1.04	2.74	1.45

0.27　−0.84　0.82　0.77　1.93　0.36　0.13

由此算出样本均值 $\bar{x}_{21}=0.91$，样本方差 $v_{21}^2=0.73$，查 t 分布表知 $t_{0.975}(20)=2.086$.

取 $d=\frac{l}{2}=0.3$，知

$$\frac{\nu_{21}^2}{d^2}(t_{0.975}(21))^2=35.29$$

从(3.3)知总样本量 τ_1 应取 36，故第二阶段应抽取样本量为 15 的样本. 设第二阶段样本如下：

1.23　2.11　1.07　1.02　1.84　−0.20　0.05　1.83

0.46　−0.06　0.69　2.32　1.05　0.44　0.30

由此算出总样本的平均值 $\bar{x}_{36}=0.92$ 及

$$\frac{\nu_{21}}{\sqrt{\tau_1}}t_{0.975}(21)=0.297$$

从(3.4)知 μ 的长度不超过 0.6 的置信区间为

$$[0.92-0.297,\ 0.92+0.297]$$

即

$$[0.623, 1.217]$$

(置信水平为 0.95).

本节的剩下部分是讨论一个特殊的估计问题——随机逼近. 随机逼近起源于 Robbins 等人(1951)的著名工作，由于应用广泛而受到人们的重视，现已形成数理统计的一个分支. 这里只讨论一些简单而典型的情况.

设 $M(x)$ 是一个函数，一个常见的问题是求方程"$M(x)=\alpha$"的根. 如果 $M(x)$ 的表达式是已知的，常用熟知的 Newton-Naphson 方法求解. 在实际问题中，往往函数 $M(x)$ 的表达式不知道，但给定 x 后，我们能得到某随机变量 $Y(x)$ 的观测值，这里 $EY(x)=M(x)$，这个 $M(x)$ 叫做回归函数. 如何找"$M(x)=\alpha$"的根或根的近似值？这是实际工作中十分关心的问题.

例 3.2 设 x 表示注入小白鼠身上的被检药物的剂量，$Y(x)$ 表示当注入 x 剂量的药物后小白鼠的反应. 将反应分成两类，一类是注入后一定时间(如 24 小时)之内死亡，另一类是这一段时间不死亡. 故

$$Y(x)=\begin{cases}1, & \text{如果小白鼠死亡}\\ 0, & \text{否则}\end{cases}$$

于是 $EY(x)\overset{\mathrm{d}}{=}M(x)=P(Y(x)=1)$ 表示平均死亡率. 所谓半数致死量 L 乃是方程"$M(x)=\frac{1}{2}$"的根，这是医学试验中很关心的量. 如何求出 L?

我们回到一般情形下如何求解的问题上来. 最早提出、也是最著名的解法是 Robbins-Monro 方法(简称 R－M 方法)，叙述如下：

取定适当的常数列 $\{a_n,\ n\geqslant 1\}$ $\left(\text{通常取 } a_n=\frac{1}{n}\right)$，先取初值 x_1 (常数)，然后用下列递推公式：

$$x_{n+1}=x_n-a_n(Y_n-\alpha)\quad(n\geqslant 1)$$

来确定序列 $\{x_n, n\geqslant 1\}$，这里 $Y_n=Y(x_n)$ 是 $Y(x)$ 在 $x=x_n$ 时的观测值$(n\geqslant 1)$. 设"$M(x)=\alpha$"的根是 θ，可以证明，在一定条件下 $P(\lim\limits_{n} x_n=\theta)=1$. 换句话说，只要 n 足够大，x_n 可作为 θ 的良好近似值.

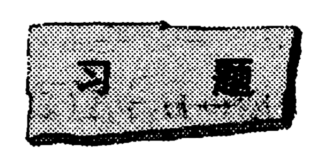
在实际问题里，有时要求回归函数 $M(x)$ 的最大值点 $\theta(M(\theta)=\max\limits_{x} M(x))$，由于 $M(x)$ 未知，不能用通常的微分法来解，又由于不能直接观测到 $M'(x)$ 的近似值，也不能使用上述的 R－M 方法. 这时可使用下列 Kiefer-Wolfowicz 方法(简称 K－W 方法，1952).

取定两个数列 $\{a_n\}$，$\{c_n\}$ $\left(\text{通常取 } a_n=\frac{1}{n},\ c_n=n^{-\frac{1}{3}},\ n\geqslant 1\right)$，先取初值 x_1(常数)，以后每一步做两个观测，用下列递推公式：

$$x_{n+1}=x_n+\frac{a_n}{c_n}(Y_{2n}-Y_{2n-1})\quad(n\geqslant 1)$$

来确定序列$\{x_n, n\geqslant 1\}$，这里$Y_{2n}=Y(x_n+c_n)$，$Y_{2n-1}=Y(x_n-c_n)$是$Y(x)$在x_n+c_n，x_n-c_n处的观测值($n\geqslant 1$)．可以证明，在一定条件下，$P(\lim_n x_n=\theta)=1$，即只要n充分大，x_n可作为$M(x)$的最大值点θ的良好近似值．

由上述可以看出，R－M方法及K－W方法均简单易行．至于保证收敛的条件及理论根据，叙述起来比较复杂，读者如有兴趣，可参看[8]．



1．设$X_1,X_2,\cdots$是独立同分布的随机变量列，且

$$P(X_1=1)=\theta=1-P(X_1=0)\quad(0<\theta\leqslant 1)$$

$$\tau=\begin{cases}\min\{n: X_n=1\}, & \text{当存在 } n\geqslant 1\text{，使得 } X_n=1\\ \infty, & \text{否则}\end{cases}$$

试证明τ是序列$\{X_n,n\geqslant 1\}$的停止法则并计算τ的数学期望．

2．甲有资本M元，乙有资本N元(M和N都是正整数)，两人进行赌博，每一局中若甲胜，则乙给甲一元；若乙胜，则甲给乙一元．设每局中甲胜的概率是$p(0<p<1)$．问：如果一局一局地赌下去，甲输光的概率是多少？平均几局后有一方输光？

3．设$X_1,X_2,\cdots$是独立同分布的随机变量列，共同分布是$N(\mu,\sigma^2)$(μ已知，σ^2未知)．给定$\sigma_1<\sigma_2$，试对检验问题

$$H_0:\sigma=\sigma_1\leftrightarrow H_1:\sigma=\sigma_2$$

求出序贯概率比检验法．

4．设$X_1,X_2\cdots$是独立同分布的随机变量列，X_1的分布密度为

$$f(x,\sigma)=\begin{cases}\dfrac{\sigma^3}{\sqrt{2\pi}}\sqrt{x}\mathrm{e}^{-\frac{\sigma^2}{2}x} & x>0\\ 0 & x\leqslant 0\end{cases}$$

其中$\sigma>0$是未知参数．

给定$0<\sigma_0<\sigma_1$，试对检验问题

$$H_0:\sigma=\sigma_0\leftrightarrow H_1:\sigma=\sigma_1$$

求出序贯概率比检验法．

5. 由于费用昂贵，在战略导弹的精度验收工作中常常至多允许发射两发导弹就要鉴定一批导弹能否通过验收. 设目标点是坐标原点，导弹的落点坐标是(X,Y)，假定X与Y相互独立，$X\sim N(0,\sigma^2)$，$Y\sim N(0,\sigma^2)$（σ未知），验收工作中的检验问题是

$$H_0:\sigma=\sigma_0\leftrightarrow H_1:\sigma=\sigma_1$$

这里$\sigma_0<\sigma_1$，σ_0是精度的设计指标. 试用类似于序贯概率比检验的思想设计出一个至多发射两发导弹的检验方案.

6. 设$X_1,X_2,\cdots$独立同分布，X_1的分布密度是$f_1(x)$或$f_2(x)$（到底是哪一个不知道），给定检验问题：

$H_1:f_1(x)$是真正的分布密度$\leftrightarrow H_2:f_2(x)$是真正的分布密度.

设(τ,d)是任一序贯检验法（其中τ是$\{X_n,\ n\geqslant1\}$的停止法则，d是判决法则），其第一类错误的概率是α，第二类错误的概率是β. 试证明$\alpha+\beta<1$时有不等式

$$E_1\tau\geqslant\frac{(1-\alpha)\ln\dfrac{\beta}{1-\alpha}+\alpha\ln\dfrac{1-\beta}{\alpha}}{E_1z}$$

$$E_2\tau\geqslant\frac{\beta\ln\dfrac{\beta}{1-\alpha}+(1-\beta)\ln\dfrac{1-\beta}{\alpha}}{E_2z}$$

这里$E_i\tau$是$f_i(x)$为真正的分布密度时τ的数学期望$(i=1,2)$，$z=\ln[f_2(X_1)/f_1(X_1)]$.

提示：利用本章引理2.1（Wald引理）.

7. 设X服从伯努利分布，X取值0或1，且

$$P_p(X=1)=p=1-P_p(X=0)$$

其中p未知，$0<p<1$，$P_p(A)$表示参数是p时事件A的概率. 给定$l\in(0,1)$，$\gamma\in(0,1)$

(1) 试证明：只要样本量足够大，p一定有置信水平为γ且长度为l的置信区间；

(2) 试用反证法证明：对任何正整数n及基于样本$X_1,X_2,\cdots X_n$的估计量$\varphi=\varphi(X_1,X_2,\cdots,X_n)$ $(0\leqslant\varphi\leqslant1)$，恒成立：

$$\inf_{0<p<1}P_p\left\{\left|\frac{\varphi-p}{p}\right|<l\right\}<\gamma$$

换句话说，在固定样本量情形下，不管样本量多么大，参数p没有相对误差足够小的估计量.

第七章　统计决策与贝叶斯统计大意

§1　统计决策问题概述

我们已经比较详细地讨论了估计和假设检验，它们可以看成是更一般的"统计决策"的特殊情形. 统计决策由四个要素组成.

设 X 的分布函数是 $F(x,\theta)$，θ 是未知参数，$\theta\in\Theta$，Θ 叫做参数空间，$\underset{\sim}{X}=(X_1,\cdots,X_n)$ 是 X 的样本. 又设 A 是某项实际工作中可能采取的各种行动所组成的非空集合，A 叫做行动空间. $L(\theta,a)$ 是定义在 $\Theta\times A$ 上的非负函数，它表示参数是 θ 时采取行动 a ($a\in A$) 引起的损失，$L(\theta,a)$ 叫做损失函数. $(\Theta,\underset{\sim}{X},A,L(\theta,a))$ 叫做统计决策问题的四个要素①.

统计决策问题是：如何根据样本 $\underset{\sim}{X}$ 的值恰当地选取行动 a 使得引起的损失尽可能的小.

下面要对这句话的确切含义进行论述.

定义 1.1　称样本空间（即样本所有可能值组成的集合）到行动空间 A 的映射 $\delta=\delta(x_1,\cdots,x_n)$ 为决策函数，简称决策.

定义 1.2　设 $\delta=\delta(x_1,\cdots,x_n)$ 是一个决策，称平均损失

$$R(\theta,\delta)=EL(\theta,\delta(X_1,\cdots,X_n))$$

为 δ 的风险.

自然想到要找风险最小的决策，但风险 $R(\theta,\delta)$ 是 θ 的函数，对一切 θ 风险最小的决策（所谓一致最优决策）难得存在.

---

① 本节是对"统计决策"作粗浅的介绍，不对有关概念给出数学上精确的陈述. 例如，不对有关的函数或映射提出"可测性"的要求.

在估计问题中，通常取 $A=\Theta$，损失函数 $L(\theta,a)$ 的类型很多，当 $A=\Theta=(\underline{\theta},\bar{\theta})$（有限或无限区间）时，最常用的有

$$L(\theta,a)=(\theta-a)^2 \tag{1.1}$$

$$L(\theta,a)=|\theta-a| \tag{1.2}$$

(1.1)叫做平方损失，(1.2)叫做绝对偏差损失．在估计问题中的决策就是估计量，在平方损失下，一致最优决策就是均方误差最小的估计量．

在假设检验问题中，参数空间 $\Theta=\Theta_0\cup\Theta_1$，$\Theta_0$ 与 Θ_1 不相交．设 H_i 是 $\theta\in\Theta_i(i=0,1)$，行动空间 $A=\{a_0,a_1\}$，其中 a_i 表示接受 $H_i(i=0,1)$．损失函数的类型很多，最常见的有

$$L(\theta,a_i)=\begin{cases}0, & \theta\in\Theta_i\\ 1, & \theta\notin\Theta_i\end{cases}\quad (i=0,1) \tag{1.3}$$

这时 $L(\theta,a)$ 叫做 0－1 损失．

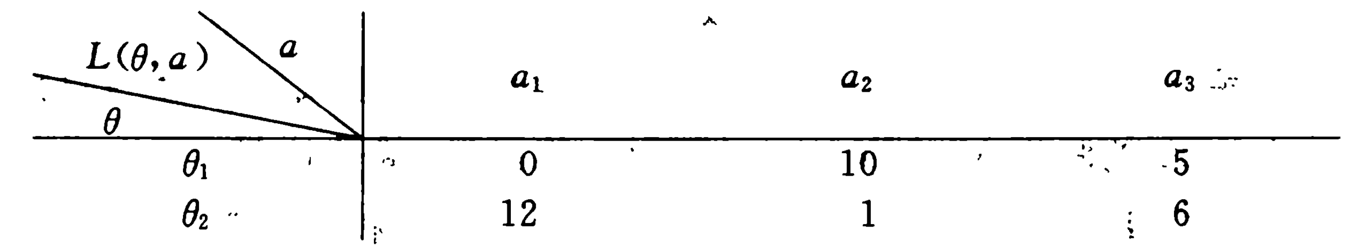
不难看出，对于假设检验问题，给出决策等价于给出对 H_0 的否定域．

例 1.1 检查某设备零件：

零件的可能状态：θ_1（好），θ_2（坏）．

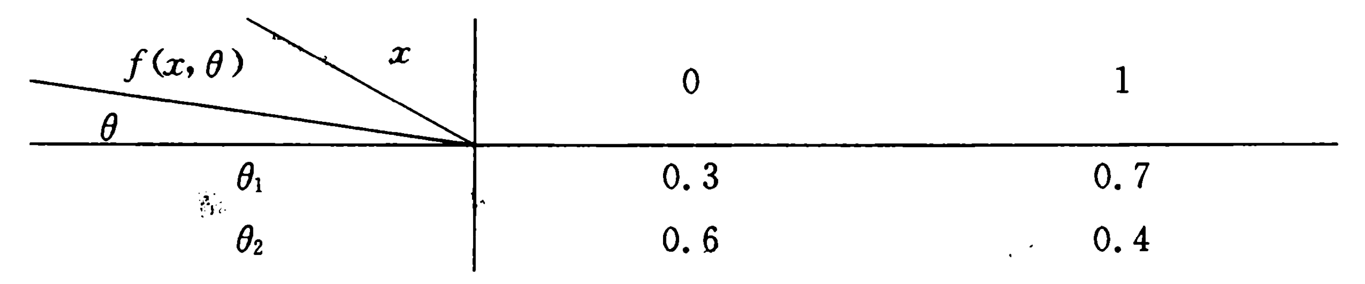
可能采取的行动：a_1（保留），a_2（更换），a_3（修理）．

损失函数为：



$L(\theta,a)$ （θ \ a）	a_1	a_2	a_3
θ_1	0	10	5
θ_2	12	1	6

样本：X 的取值为 0 或 1（为判断零件的状态，在设备工作时用手摸零件，温度正常时则记 $X=1$，发烫则记 $X=0$）．X 的概率函数为



$f(x,\theta)$ （θ \ x）	0	1
θ_1	0.3	0.7
θ_2	0.6	0.4

问:应采取怎样的决策使得损失尽可能小?

可能的决策函数有 9 个,列表如下:

X \ δ	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
0	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
1	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3

风险为

$$
\begin{aligned}
R(\theta,\delta) &= EL(\theta,\delta(X)) \\
&= L(\theta,a_1)P(\delta(X)=a_1)+L(\theta,a_2)P(\delta(X)=a_2)+ \\
&\quad L(\theta,a_3)P(\delta(X)=a_3)
\end{aligned}
$$

例如

$$R(\theta_1,\delta_2)=0\times 0.3+10\times 0.7+5\times 0=7$$

$$R(\theta_2,\delta_2)=12\times 0.6+1\times 0.4+6\times 0=7.6$$

9 个决策的风险如下

δ	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
$R(\theta_1,\delta)$	0	7	3.5	3	10	6.5	1.5	8.5	5
$R(\theta_2,\delta)$	12	7.6	9.6	5.4	1	3	8.4	4.0	6

哪个决策好呢? 有些是明显的,如 δ_4 优于 $\delta_2,\delta_3,\delta_9$. 但 δ_4 与 δ_1 哪个好呢? 这就不好回答了.

定义 1.3 称决策 $\delta=\delta(x_1,\cdots,x_n)$ 是容许的,如果不存在另一决策 δ^1 使得

$$R(\theta,\delta^1)\leqslant R(\theta,\delta) \quad (\text{一切 } \theta\in\Theta)$$

且对至少一个 θ,严格的不等式成立.

不难看出,在前面的例 1.1 中,$\delta_1,\delta_4,\delta_5,\delta_6,\delta_7$ 是容许的,δ_2,δ_3,δ_9 是不容许的,而且不存在一致最优的决策.

如果一致最优决策不存在,只好去找比较“优良”的决策了. 有两个途径探讨这个问题. 一是对决策 δ 作一定的限制,缩小选择的范围,在较小的范围内找最优的,例如在估计问题中要求 δ 是无偏的估计,在所有无偏估计中找最优的. 另一是放宽比较的要

求，不要求风险函数对一切 θ 都最小.

定义 1.4 称决策 δ^* 是 minimax 决策，若对一切决策 δ 成立

$$\sup_\theta R(\theta,\delta^*)\leqslant\sup_\theta R(\theta,\delta)$$

minimax 决策是一种保守的决策. 不难看出，在例 1.1 中 δ_4 是 minimax 决策.

当然，在许多情形下应尽量避免采用保守的决策. 为此，应对参数 θ 有所了解，知道哪些 θ 值出现机会较多，哪些 θ 值出现机会较少. 换句话说，应将 θ 看成一个随机变量，它有确定的概率分布. 在这种情形下，风险 $R(\theta,\delta)$ 关于 θ 的分布的平均值 $\rho(\delta)=ER(\theta,\delta)$ 是评价决策 δ 的优良性指标，数值越小越好. 把未知参数 θ 看成是随机变量，这是一种重要的观点，对统计学有重大影响，下面要作进一步介绍.

§2 什么是贝叶斯统计

公元 1763 年，贝叶斯（T. Bayes，1702—1761，英国长老会的牧师）的著作《论有关机遇问题的求解》发表了。这篇著作提出了一种归纳推理的理论。这个理论以后被一些学者发展为一种系统的统计推断方法，称为贝叶斯方法. 采用这种方法作统计推断所得的全部结果，构成贝叶斯统计的内容. 认为贝叶斯方法是唯一合理的统计推断方法的统计学者，组成统计学中的贝叶斯学派. 这个学派始自 20 世纪 30 年代，到 50、60 年代引起人们广泛的注意. 时至今日，其影响有日益扩大之势.

设总体 X 有分布密度（或概率函数）$f(x,\theta)$，其中 θ 是未知的，但知 θ 的变化范围是 Θ，经典方法把参数 θ 看作是客观常数，通过对样本 $\underset{\sim}{X}=(X_1,\cdots,X_n)$ 的研究对 θ 给出估计值或者推断 θ 属于某个给定的范围. 贝叶斯学派的根本观点，是认为在关于 θ 的任何统计推断问题中，除了使用样本 $\underset{\sim}{X}$ 提供的信息外，还必须

对 θ 规定一个先验分布，它是进行推断时不可缺少的要素. 说得更明确些，应把 θ 看成随机变量（为了强调这一点，有时把 θ 写成 $\tilde{\theta}$），它服从某个概率分布（叫做先验分布），总体 X 的分布实际上是 θ 给定时 X 的条件分布. 贝叶斯学派把先验分布看成是在抽样（或观测）前就有的关于 θ 的先验信息的概率表述，先验分布可能有客观的依据，也可以部分地或是全部地基于主观信念. 根据 X 的分布密度（或概率函数）$f(x,\theta)$ 及 θ 的先验分布密度（或先验概率函数）$\xi(\theta)$，可以算出在样本 $\underset{\sim}{X}=(X_1,\cdots,X_n)=\underset{\sim}{x}=(x_1,\cdots,x_n)$ 的条件下，θ 的条件分布密度（或条件概率函数）$\xi(\theta|\underset{\sim}{x})$. 因为这个分布是在抽样（或观测）之后得到的，故称为后验分布. 贝叶斯学派认为，这个分布综合了样本 $\underset{\sim}{X}$ 及先验分布密度（或概率函数）$\xi(\theta)$ 所提供的信息. 抽样（或观测）的全部目的就在于完成由先验分布到后验分布的转换. 贝叶斯方法的关键在于所作出的任何推断都只须根据后验分布 $\xi(\theta|\underset{\sim}{x})$，而不再涉及样本 $\underset{\sim}{X}$ 的分布. 但在如何使用 $\xi(\theta|\underset{\sim}{x})$ 上还有较大的灵活性，涉及到行动空间和损失函数的类型.

沿用§1中的记号，用 A 表示行动空间，用 $L(\theta,a)$ 表示损失函数. 若参数 θ 的先验分布是 ξ，决策是 $\delta=\delta(x_1,\cdots,x_n)$，则平均风险为 $\rho(\delta)=E_\xi R(\theta,\delta)$，这里 $R(\theta,\delta)$ 是风险，E_ξ 是关于 θ 的概率分布 ξ 取平均值.

定义 2.1 称 $\delta^*=\delta^*(x_1,\cdots,x_n)$ 是贝叶斯决策，若

$$\rho(\delta^*)=\inf_{\delta}\rho(\delta)$$

此时 $\rho(\delta^*)$ 叫做贝叶斯风险.

要注意的是，贝叶斯决策依赖于先验分布. 先验分布变了，贝叶斯决策一般也要变.

例 2.1 在例 1.1 中，若 θ 的先验分布（概率函数）是 $\xi(\theta_1)=0.7$，$\xi(\theta_2)=0.3$，则

$$\rho(\delta)=0.7\times R(\theta_1,\delta)+0.3\times R(\theta_2,\delta)$$

各个决策的 $\rho(\delta)$ 值如下：

δ	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
$\rho(\delta)$	3.6	5.13	12.48	3.72	7.20	5.45	3.57	7.15	5.3

从上表可看出

$$\delta_7:\begin{matrix}\text{发烫}(X=0)\text{，就修理}(a_3)\\ \text{正常}(X=1)\text{，就保留}(a_1)\end{matrix}$$

是贝叶斯决策.

若先验分布改为 $\xi(\theta_1)=0.9$，$\xi(\theta_2)=0.1$，则贝叶斯决策是 δ_1.

贝叶斯统计的基本任务就是针对先验分布 $\xi(\theta)$，找出贝叶斯决策.

怎样寻找贝叶斯决策呢？例 2.1 所处理的是很简单的情形，穷举所有的决策，加以比较后找出贝叶斯决策. 一般情形下不能这样做，而要利用后验分布. 设 $\xi(\theta|x_1,\cdots,x_n)$ 是样本 $(X_1,\cdots,X_n)=(x_1,\cdots,x_n)$ 下 θ 的条件分布密度(后验分布①). 令

$$\rho(x_1,\cdots,x_n)=\inf_{a\in A}\int_\Theta L(\theta,a)\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta \qquad (2.1)$$

这个 $\rho(x_1,\cdots,x_n)$ 乃是样本值为 $x_1,\cdots,x_n$ 时的最小后验平均损失.

若决策 $\delta^*=\delta^*(x_1,\cdots,x_n)$ 满足

$$\int_\Theta L(\theta,\delta^*(x_1,\cdots,x_n))\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta=\rho(x_1,\cdots,x_n) \qquad (2.2)$$

$$(\text{对一切 } x_1,\cdots,x_n)$$

则 δ^* 就是贝叶斯决策. 这个结论的数学证明用到较深的数学知识，从略. 许多书上干脆把满足(2.2)的 δ^* 定义为贝叶斯决策. (2.2)的意义是：采用决策 δ^* 引起的后验平均损失达到最小值.

(2.2)是寻找贝叶斯决策的基本出发点. 为此首先需要求出后验分布.

怎样计算后验分布呢？可以证明，当 θ 的先验分布密度是 $\xi(\theta)$ 时，在 $\underset{\sim}{X}=(x_1,\cdots,x_n)$ 的条件下 θ 的后验分布密度是

① 当 θ 是离散型随机变量时，可进行类似的讨论，下面公式里的积分要改为求和.

$$\xi(\theta|x_1,\cdots,x_n)=\frac{\prod_{i=1}^{n}f(x_i,\theta)\xi(\theta)}{\int_{\Theta}\prod_{i=1}^{n}f(x_i,\theta)\xi(\theta)\mathrm{d}\theta} \tag{2.3}$$

当 $\Theta=\{\theta_1,\theta_2,\cdots\}$时，(2,3)要用下式代替

$$\xi(\theta|x_1,\cdots,x_n)=\frac{\prod_{1}^{n}f(x_i,\theta)\xi(\theta)}{\sum_{j}\prod_{1}^{n}f(x_i,\theta_j)\xi(\theta_j)} \tag{2.4}$$

这时(2.4)是 θ 的后验概率函数.

(2.3)的严格证明涉及到零概率事件发生下如何计算条件概率的问题，要用测度论才能处理①，当 X 及 θ 是离散型随机变量时，

---

① 虽然本讲义不给出(2.3)的严格证明，但我们可给出下列有道理的说明：

$$P(\tilde{\theta}\leqslant b\,|\,x_i-\varepsilon\leqslant X_i\leqslant x_i+\varepsilon,i=1,2,\cdots,n)$$

$$=\frac{P(\tilde{\theta}\leqslant b,x_i-\varepsilon\leqslant X_i\leqslant x_i+\varepsilon,i=1,2,\cdots,n)}{P(x_i-\varepsilon\leqslant X_i\leqslant x_i+\varepsilon,i=1,2,\cdots,n)}$$

但 $\theta,X_1,\cdots,X_n$ 的联合分布密度是 $\xi(\theta)\prod_{i=1}^{n}f(x_i,\theta)$，故

$$P(\tilde{\theta}\leqslant b\,|\,x_i-\varepsilon\leqslant X_i\leqslant x_i+\varepsilon,i=1,2,\cdots,n)$$

$$=\frac{\frac{1}{(2s)^n}\int_{\underline{\theta}}^{b}\int_{x_1-\varepsilon}^{x_1+\varepsilon}\cdots\int_{x_n-\varepsilon}^{x_n+\varepsilon}\xi(\theta)\prod_{1}^{n}f(u_i,\theta)\mathrm{d}\theta\mathrm{d}u_1\cdots\mathrm{d}u_n}{\frac{1}{(2\varepsilon)^n}\int_{\underline{\theta}}^{\bar{\theta}}\int_{x_1-\varepsilon}^{x_1+\varepsilon}\cdots\int_{x_n+\varepsilon}^{x_n+\varepsilon}\xi(\theta)\prod_{1}^{n}f(u_i,\theta)\mathrm{d}\theta\mathrm{d}u_1\cdots\mathrm{d}u_n}$$

这里$(\underline{\theta},\bar{\theta})=\Theta$. 当 $f(x,\theta)$ 是 x 的连续函数时，知

$$\lim_{\varepsilon\to 0}P(\tilde{\theta}\leqslant b\,|\,x_i-\varepsilon\leqslant X_i\leqslant x_i+\varepsilon,i=1,2,\cdots,n)$$

$$=\frac{\int_{\underline{\theta}}^{b}\xi(\theta)\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\theta}{\int_{\underline{\theta}}^{\bar{\theta}}\xi(\theta)\prod_{1}^{n}f(x_i,\theta)\mathrm{d}\theta}$$

从直观上看，上式右边应该就是 $P(\tilde{\theta}\leqslant b\,|\,X_1=x_1,\cdots,X_n=x_n)$. 由于 b 的任意性，知 θ 的后验分布密度是

$$\frac{\xi(\theta)\prod_{1}^{n}f(x_i,\theta)}{\int_{\Theta}\xi(\theta)\prod_{1}^{n}f(x_i,\theta)}$$

(2.4)的证明是不难的. 实际上这时 $\xi(\theta_j)=P(\tilde{\theta}=\theta_j)$, $f(x_i;\theta_j)=P(X=x_i|\tilde{\theta}=\theta_j)$, 于是

$$P(\tilde{\theta}=\theta_j|X_1=x_1,\cdots,X_n=x_n)$$

$$=\frac{P(\tilde{\theta}=\theta_j,X_1=x_1,\cdots,X_n=x_n)}{P(X_1=x_1,\cdots,X_n=x_n)}$$

$$=\frac{P(\tilde{\theta}=\theta_j)\prod_1^n P(X_i=x_i|\tilde{\theta}=\theta_j)}{\sum_j P(\tilde{\theta}=\theta_j,X_1=x_1,\cdots,X_n=x_n)}$$

$$=\frac{\xi(\theta_j)\prod_1^n f(x_i,\theta_j)}{\sum_j\prod_1^n f(x_i,\theta_j)\xi(\theta_j)}$$

这就证明了(2.4). (2.4)就是初等概率论中的贝叶斯公式.

例 2.2 设 X 服从伯努利分布,参数是 p,即

$$P(X=1)=p=1-P(X=0);\quad 0<p<1$$

设 p 的先验分布 ξ 是(0,1)上的均匀分布. 从(2.1)知在样本 $X_1=x_1,\cdots,X_n=x_n$ 下 p 的后验分布密度是

$$\xi(p|x_1,\cdots,x_n)=\frac{p^{\sum_1^n x_i}\cdot(1-p)^{n-\sum_1^n x_i}}{\int_0^1 p^{\sum_1^n x_i}\cdot(1-p)^{n-\sum_1^n x_i}\mathrm{d}p}\tag{2.5}$$

这是参数为 $\sum_1^n x_i+1,n+1-\sum_1^n x_i$ 的 β 分布.

设行动空间 $A=(0,1)$,损失函数是(1.1)(平方损失),则(2.2)式化为

$$\int_0^1(p-\delta^*(x_1,\cdots,x_n))^2\xi(p|x_1,\cdots,x_n)\mathrm{d}p$$

$$=\inf_{a\in(0,1)}\int_0^1 (p-a)^2\xi(p|x_1,\cdots,x_n)\mathrm{d}p \tag{2.6}$$

从(2.6)式出发可以证明：$\delta^*(x_1,\cdots,x_n)$恰好是后验分布$\xi(p|x_1,\cdots,x_n)$的均值. 利用(2.5)知

$$\delta^*(x_1,\cdots,x_n)=\int_0^1 p\xi(p|x_1,\cdots,x_n)\mathrm{d}p$$

$$=\frac{1}{n+2}\left(\sum_1^n x_i+1\right)$$

这就是 p 的贝叶斯估计.

例 2.3 设随机变量 X 的密度函数是 $f(x,\theta)$，$\theta\in\Theta$，$\Theta=\Theta_0\cup\Theta_1$，$\Theta_1$ 与 Θ_0 不相交，θ 的先验分布密度是 $\xi(\theta)$. 检验问题是：

$$H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1$$

设 $A=\{a_0,a_1\}$（a_i 表示接受假设 H_i，$i=0,1$）. 若 X 的样本$(X_1,\cdots,X_n)$取值$(x_1,\cdots,x_n)$，θ 的后验分布密度为 $\xi(\theta|x_1,\cdots,x_n)$，又 $L(\theta,a)$是 0－1 损失（见(1.3)）. 则

$$\int_\Theta L(\theta,a_i)\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta$$

$$=P(\tilde{\theta}\notin\Theta_i|X_1=x_1,\cdots,X_n=x_n)$$

$$=1-P(\tilde{\theta}\in\Theta_i|X_1=x_1,\cdots,X_n=x_n)\quad(i=0,1)$$

从(2.2)知贝叶斯决策为

$$\delta^*(x_1,\cdots,x_n)=\begin{cases}a_0, & 当\ P(\tilde{\theta}\in\Theta_0|X_1=x_1,\cdots,X_n=x_n)\geqslant\frac{1}{2}\\ a_1, & 否则\end{cases}$$

换句话说，贝叶斯检验是：当且仅当条件概率

$$P(\tilde{\theta}\in\Theta_0|X_1=x_1,\cdots,X_n=x_n)$$

不小于$\frac{1}{2}$时接受假设 H_0.

对于一般的决策问题（包含估计、假设检验在内），可以证明在相当广泛的条件下，贝叶斯决策是存在的，读者可参看[8].

作为本节的末尾，我们还要简略地介绍贝叶斯序贯统计决策.

设 X 有分布密度(或概率函数)$f(x,\theta)$,θ 取值属于 Θ,有先验分布 $\xi(\theta)$,A 是行动空间,$L(\theta,a)$ 是损失函数. 设对 X 进行逐次观测,得 $X_1,X_2,\cdots$,单次观测的费用是 $C>0$.

序贯统计决策由停止法则 τ 及判决法则 δ 组成. 停止法则 τ 告诉我们何时停止观测,τ 是不依赖于将来的随机变量(见第六章);判决法则 δ 告诉我们,如何根据序贯样本 $X_1,\cdots,X_\tau$ 选取行动 $\delta(X_1,\cdots,X_\tau)$($\delta$ 的值属于 A).

序贯统计决策(τ,δ)的总风险为

$$\rho(\tau,\delta)=E[L(\theta,\delta(X_1,\cdots,X_\tau))+\tau C]$$

(损失和费用之和的平均值). 这里 E 表示 θ 之先验分布为 ξ 及给定 θ 时诸 X_i 的分布密度(或概率函数)为 $f(x,\theta)$时计算期望.

定义 2.2 称(τ^*,δ^*)是贝叶斯序贯决策,若

$$\rho(\tau^*,\delta^*)=\inf_{(\tau,\delta)}\rho(\tau,\delta)$$

设 θ 的先验分布密度是 $\xi(\theta)$,$X_1=x_1,\cdots,X_n=x_n$ 下 θ 的后验分布密度为 $\xi(\theta|x_1,\cdots,x_n)$. 又

$$\{\tau=n\}=\{(X_1,\cdots,X_n)\in B_n\}$$

可以证明

$$\rho(\tau,\delta)=\sum_n\int_{B_n}\left\{\int_\Theta[L(\theta,\delta(x_1,\cdots,x_n))+nC]\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta\right\}$$
$$f_n(x_1,\cdots,x_n)\mathrm{d}x_1,\cdots,\mathrm{d}x_n \tag{2.7}$$

这里 $f_n(x_1,\cdots,x_n)$是 $X_1,\cdots,X_n$ 的联合密度,

$$f_n(x_1,\cdots,x_n)=\int_\Theta\prod_{i=1}^n f(x_i,\theta)\xi(\theta)\mathrm{d}\theta$$

若 δ^* 是这样的:对一切 $a\in A$ 及 $n\geqslant 1$ 有

$$\int_\Theta L(\theta,a)\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta$$
$$\geqslant\int_\Theta L(\theta,\delta^*(x_1,\cdots,x_n))\xi(\theta|x_1,\cdots,x_n)\mathrm{d}\theta$$
$$(\text{一切 } x_1,\cdots,x_n) \tag{2.8}$$

从(2.7)不难看出

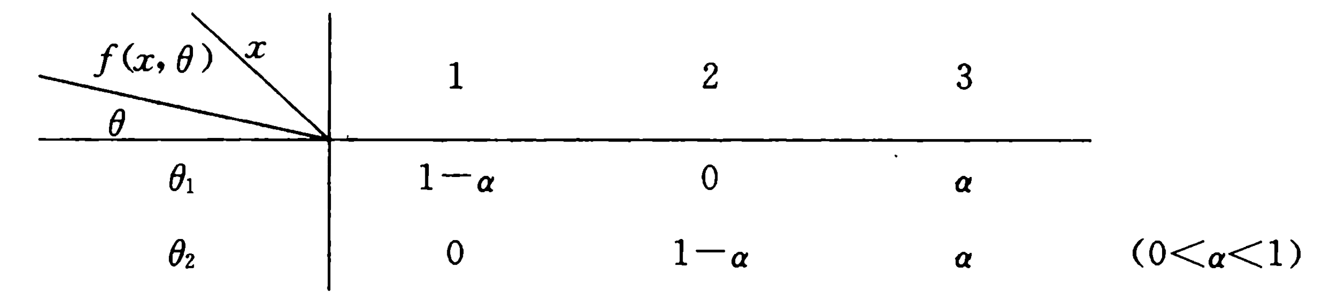
$$\rho(\tau,\delta)\geqslant\rho(\tau,\delta^*)$$

为了使总风险达到最小,只须找停止法则 τ^* 使得

$$\rho(\tau^*,\delta^*)=\inf_{\tau}\rho(\tau,\delta^*)$$

这就化成了著名的最优停止问题. 现代已有丰富的最优停止理论,读者可参阅[8].

例 2.4 设 $\Theta=\{\theta_1,\theta_2\}$, $A=\{a_1,a_2\}$, X 的取值空间为$\{1,2,3\}$, X 的概率函数 $f(x,\theta)$是这样的

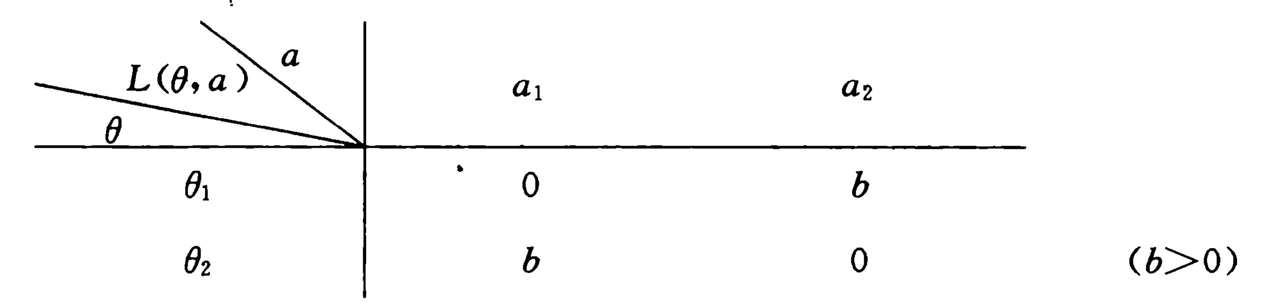


$f(x,\theta)$ (θ \ x)	1	2	3	
θ_1	$1-\alpha$	0	α	
θ_2	0	$1-\alpha$	α	$(0<\alpha<1)$

设 θ 的先验分布为

$$P(\tilde{\theta}=\theta_1)=\xi=1-P(\tilde{\theta}=\theta_2)\quad\left(0<\xi\leqslant\frac{1}{2}\right)$$

损失函数 $L(\theta,a)$是这样的:



$L(\theta,a)$ (θ \ a)	a_1	a_2	
θ_1	0	b	
θ_2	b	0	$(b>0)$

又设单次观测的费用是 $C>0$. 我们指出,存在序贯统计决策比固定样本量的决策要好.

首先求出后验分布. 不难看出,

$$P(X_1=x_1,\cdots,X_n=x_n)$$
$$=\xi\prod_{i=1}^{n}f(x_i,\theta_1)+(1-\xi)\prod_{i=1}^{n}f(x_i,\theta_2)$$

可见,若 $x_1,\cdots,x_n$ 中既含有 1,又含有 2,则上述概率为 0. 以下不妨设为 $x_1,\cdots,x_n$ 中不同时含有 1 和 2. 经计算知

$$\xi(\theta_1 \mid x_1, \cdots, x_n) = P(\tilde{\theta} = \theta_1 \mid X_1 = x_1, \cdots, X_n = x_n)$$

$$= \begin{cases} 1, & \text{当 } x_1, \cdots, x_n \text{ 中含有 } 1 \\ 0, & \text{当 } x_1, \cdots, x_n \text{ 中含有 } 2 \\ \xi, & \text{当 } x_1 = \cdots = x_n = 3 \end{cases}$$

$$\xi(\theta_2 \mid x_1, \cdots, x_n) = 1 - \xi(\theta_1 \mid x_1, \cdots, x_n)$$

易知,最小后验平均损失

$$\rho(x_1, \cdots, x_n) \overset{\mathrm{d}}{=} \min_a \sum_{i=1}^{2} L(\theta_i, a)\xi(\theta_i \mid x_1, \cdots, x_n)$$

$$= b\min(\xi(\theta_1 \mid x_1, \cdots, x_n), \xi(\theta_2 \mid x_1, \cdots, x_n))$$

$$= \begin{cases} b\xi, & \text{当 } x_1 = x_2 = \cdots = x_n = 3 \\ 0, & \text{否则} \end{cases}$$

令

$$\delta_n^*(x_1, \cdots, x_n) = \begin{cases} a_1, & \text{当 } x_1, \cdots, x_n \text{ 中含有 } 1 \\ a_2, & \text{否则} \end{cases} \tag{2.9}$$

易知,$\delta_n^*(x_1, \cdots, x_n)$的后验平均损失正好是$\rho(x_1, \cdots, x_n)$. 由此可见,对任何固定的样本量 n,贝叶斯决策是 δ_n^*. 此时的总风险(损失与观测费用之和的平均值)为

$$\rho(n) = E\rho(X_1, \cdots, X_n) + nC$$

$$= b\xi P(X_1 = \cdots = X_n = 3) + nC$$

$$= b\xi\alpha^n + nC$$

自然问:n 取何值时,$\rho(n)$最小?

设

$$\frac{\xi b}{C} > \frac{1}{\alpha \cdot |\ln \alpha|} \tag{2.10}$$

易知 $\xi b\alpha^x + Cx$ 在

$$x = n^* = \frac{1}{|\ln \alpha|} \ln\left(\frac{\xi b |\ln \alpha|}{C}\right)$$

时达最小值:

$$\rho(n^*)=\frac{C}{|\ln\alpha|}\left(1+\ln\frac{\xi b|\ln\alpha|}{C}\right)$$

(这里假定 n^* 是整数).

以上讨论表明,所有固定样本量的贝叶斯决策中取样本量为 n^* 最好.

令

$$N=\inf\{n: n\geqslant 1\ X_n\neq 3\}$$

这 N 是一个停止法则. 用 δ^* 表示这样的判决法则,对一切 $x_1,\cdots,x_n$, $\delta^*(x_1,\cdots,x_n)=\delta_n^*(x_1,\cdots,x_n)$($\delta_n^*$ 的定义见(2.9)).

我们指出,序贯统计决策(N,δ^*)的总风险小于 $\rho(n^*)$.

实际上,此时

$$\begin{aligned}\{N=n\}&=\{X_1=3,\cdots,X_{n-1}=3,X_n\neq 3\}\\&=\{(X_1,\cdots,X_n)\in B_n\}\end{aligned}$$

其中

$$B_n=\{(\underbrace{3,\cdots,3}_{n-1\text{个}},1),(\underbrace{3,\cdots,3}_{n-1\text{个}},2)\}$$

从(2.7)知

$$\begin{aligned}\rho(N,\delta^*)&=CEN\\&=C\sum_{n=1}^{\infty}nP(N=n)\\&=C\sum_{n=1}^{\infty}nP(X_1=3,\cdots,X_{n-1}=3,X_n\neq 3)\\&=C\sum_{n=1}^{\infty}n\alpha^{n-1}(1-\alpha)\\&=\frac{C}{1-\alpha}\end{aligned}$$

另一方面,从(2.10)知

$$\rho(n^*)>\frac{C}{|\ln\alpha|}(1+|\ln\alpha|)>\frac{C}{1-\alpha}$$

故

$$\rho(N,\delta^*)<\rho(n^*)$$

这表明，(N,δ^*)的总风险小于任何固定样本量方法的总风险.

§3 先验分布的确定

如何确定先验分布，这是贝叶斯统计中最困难，也是使用贝叶斯方法必须解决但又最易引起争议的问题. 这个问题现代有很多研究成果，但还没有圆满的理论与普遍有效的方法（参看 S. Kotz 与吴喜之合著的《现代贝叶斯统计学》，中国统计出版社，2000 及 J. O. Berger 著的《统计决策论及贝叶斯分析》（贾乃光译），中国统计出版社，1998）. 从实用角度出发，应充分利用专家的经验或者对历史上积累的数据进行分析和拟合，以确定先验分布. 在确定先验分布时，许多人利用下列的协调性假说.

协调性假说：若总体 X 的分布密度（或概率函数）是 $f(x,\theta)$，则 θ 的先验分布与由它和 X 的样本确定的后验分布应属于同一类型. 这时先验分布叫做是与 $f(x,\theta)$ 共轭的分布.

这里未对"同一类型"四字给出精确的定义，也很难给出恰当的定义. 通常的理解是，将概率性质相似的所有分布算作同一类型. 例如，所有正态分布构成一类；所有 Γ 分布构成一类；所有 β 分布构成一类.

这个假说指示我们，先验分布应取何种类型，然后再利用历史数据来确定先验分布中的未知部分. 许多实践表明，这个假说是符合实际的.

以下我们要对一些常见的分布找出其共轭的先验分布.

定理 3.1 设 $X_1,\cdots,X_n$ 是来自伯努利分布（参数是 p，$0<p<1$）的样本. 若 p 的先验分布是 β 分布，参数是 α,β，则在 $X_1=x_1,\cdots,X_n=x_n$ 下 p 的后验分布是参数为 $\alpha+y,\beta+n-y$ 的 β 分布，这里 $y=\sum_1^n x_i$.

证明：$X_1,\cdots,X_n$ 的联合概率函数为

$$f_n(x_1,\cdots,x_n;p)=\begin{pmatrix} n \\ \sum_1^n x_i \end{pmatrix} p^{\sum_1^n x_i}\cdot(1-p)^{n-\sum_1^n x_i}$$

p 的先验分布密度

$$\xi(p)=C(\alpha,\beta)p^{\alpha-1}(1-p)^{\beta-1}$$

（这里 $C(\alpha、\beta)$ 是常数）. 从公式(2.3)知 p 的后验分布密度为

$$\xi(p\,|\,x_1,\cdots,x_n)=\frac{f_n(x_1,\cdots,x_n;p)\xi(p)}{\int_0^1 f_n(x_1,\cdots,x_n;p)\xi(p)\mathrm{d}p}$$

$$=Ap^{\alpha+\sum_1^n x_i-1}\cdot(1-p)^{\beta+n-\sum_1^n x_i-1}$$

这里 A 是与 p 无关的数，可见后验分布是参数为 $\alpha+y,\beta+n-y$ 的 β 分布. 证毕.

定理 3.1 告诉我们，β 分布族是伯努利分布族的共轭分布族.

定理 3.2 设 $X_1,\cdots,X_n$ 是来自 Poisson 分布的样本. Poisson 分布的参数是 λ，λ 的先验分布是 Γ 分布（参数是 α,β），则在 $X_1=x_1,\cdots,X_n=x_n$ 下 λ 的后验分布是参数为 $\alpha+\sum_1^n x_i,\beta+n$ 的 Γ 分布.

证明：$X_1,\cdots,X_n$ 的联合概率函数为

$$f_n(x_1,\cdots,x_n;\lambda)=\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}\mathrm{e}^{-n\lambda}$$

λ 的先验分布密度是

$$\xi(\lambda)=C\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}\quad(\lambda>0)$$

根据(2.3)知 λ 的后验分布密度为

$$\xi(\lambda\,|\,x_1,\cdots,x_n)=A\lambda^{\alpha+\sum_1^n x_i-1}\cdot\mathrm{e}^{-(\beta+n)\lambda}\quad(\lambda>0)$$

其中 A 是与 λ 无关的常数. 这表明后验分布是参数为 $\alpha+\sum_1^n x_i$，$\beta+n$ 的 Γ 分布. 证毕.

定理 3.3 设 $X_1,\cdots,X_n$ 是来自指数分布 $F(x)=1-e^{-\lambda x}$ ($\lambda>0,x>0$)的样本. 设参数 λ 的先验分布是参数为 α,β 的 Γ 分布，则在 $X_1=x_1,\cdots,X_n=x_n$ 下 λ 的后验分布是参数为 $\alpha+n,\beta+\sum_1^n x_i$ 的 Γ 分布.

证明：$X_1,\cdots,X_n$ 的联合密度为

$$f_n(x_1,\cdots,x_n;\lambda)=\lambda^n e^{-\lambda\sum_1^n x_i}$$

λ 的先验分布密度为

$$\xi(\lambda)=C\lambda^{\alpha-1}e^{-\beta\lambda}\quad(\lambda>0)$$

从公式(2.3)知 λ 的后验分布密度

$$\xi(\lambda|x_1,\cdots,x_n)=A\lambda^{\alpha+n-1}\cdot e^{-(\beta+\sum_1^n x_i)\lambda}$$

可见后验分布是参数为 $\alpha+n,\beta+\sum_1^n x_i$ 的 Γ 分布.　　证毕.

定理 3.4 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的样本，其中 σ^2 已知，μ 未知，设 μ 的先验分布为 $N(\mu_0,\sigma_0^2)$，则在 $X_1=x_1,\cdots,X_n=x_n$ 下，μ 的后验分布为 $N(\mu^*,(\sigma^*)^2)$，其中

$$\mu^*=\frac{\mu_0\sigma^2+n\sigma_0^2(\bar{x})^2}{\sigma^2+n\sigma_0^2}\quad\left(\bar{x}=\frac{1}{n}\sum_1^n x_i\right)\tag{3.1}$$

$$(\sigma^*)^2=\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\tag{3.2}$$

证明：$X_1,\cdots,X_n$ 的联合密度为

$$f_n(x_1,\cdots,x_n;\mu)=C\exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right\}$$

这里 C 与 μ 无关. 又 μ 的先验分布密度为

$$\xi(\mu)=C_1\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

从公式(2.3)知后验分布密度为

$$\xi(\mu|x_1,\cdots,x_n)=A\exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

其中 A 与 μ 无关. 但是

$$
\begin{aligned}
&\frac{n}{\sigma^2}(\mu-\bar{x})^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2 \\
&=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\mu\left(\frac{n(\bar{x})^2}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)+\frac{n}{\sigma^2}(\bar{x})^2+\frac{\mu_0^2}{\sigma_0^2} \\
&=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)(\mu-\mu^*)^2+B
\end{aligned}
$$

其中 μ^* 由(3.1)确定,B 与 μ 无关. 于是

$$
\xi(\mu\,|\,x_1,\cdots,x_n)=\widetilde{A}\exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)(\mu-\mu^*)^2\right\}
$$

可见后验分布是 $N(\mu^*,(\sigma^*)^2)$. 证毕.

从定理 3.4 看出,μ^* 是 $\bar{x}$ 与 μ_0 的加权平均. 这与直观相符,σ_0^2 越大,μ_0 的值越不重要.

定理 3.5 设 $X_1,\cdots,X_n$ 是来自 $N\left(\mu,\frac{1}{R}\right)$ 的样本,μ 已知,R 未知,设 R 的先验分布是参数为 α,β 的 Γ 分布. 则在 $X_1=x_1,\cdots,X_n=x_n$ 下,R 的后验分布是参数为 α',β' 的 Γ 分布,这里 $\alpha'=\alpha+\frac{n}{2},\beta'=\beta+\frac{1}{2}\sum\limits_{1}^{n}(x_i-\mu)^2$.

证明:$X_1,\cdots,X_n$ 的联合密度为

$$
f_n(x_1,\cdots x_n;R)=C_1R^{\frac{n}{2}}\ \mathrm{e}^{-\frac{1}{2}R\sum\limits_1^n(x_i-\mu)^2}
$$

C_1 是与 R 无关的常数. R 的先验分布密度为

$$
\xi(R)=C_2R^{\alpha-1}\mathrm{e}^{-\beta R}
$$

从公式(2.3)知 R 的后验分布密度为

$$
CR^{\alpha+\frac{n}{2}-1}\cdot\mathrm{e}^{-\left(\beta+\frac{1}{2}\sum\limits_1^n(x_i-\mu)^2\right)R}
$$

可见后验分布是参数为 α',β' 的 Γ 分布. 证毕.

为了研究均匀分布的共轭分布,引进定义 3.1.

定义 3.1 称 X 服从参数为 $x_0>0,\alpha>0$ 的 Pareto 分布,若它的分布密度是

$$\varphi(x;x_0,\alpha)=\begin{cases}0, & x<x_0\\ \alpha x_0^{\alpha}x^{-\alpha-1}, & x\geqslant x_0\end{cases}$$

定理 3.6 设 $X_1,\cdots,X_n$ 是来自$[0,\theta]$上均匀分布的样本. 若 θ 的先验分布是参数为 θ_0,α 的 Pareto 分布. 则在 $X_1=x_1,\cdots,X_n=x_n$ 下 θ 的后验分布是参数为$\theta_0',\alpha+n$ 的 Pareto 分布,这里 $\theta_0'=\max\{\theta_0,x_1,\cdots,x_n\}$.

证明:$X_1,\cdots,X_n$ 的联合密度是

$$f_n(x_1,\cdots,x_n;\theta)=\frac{1}{\theta^n}\prod_{i=1}^{n}I_{[0,\theta]}(x_i)$$

$$=\frac{1}{\theta^n}I_{[0,\theta]}(\max_{1\leqslant i\leqslant n}x_i)$$

而 θ 的先验分布密度是

$$\varphi(\theta;\theta_0,\alpha)=\frac{\alpha\theta_0^{\alpha}}{\theta^{\alpha+1}}I_{[\theta_0,\infty)}(\theta)$$

根据(2.3)知 θ 的后验分布密度是

$$\xi(\theta|x_1,\cdots,x_n)=A\theta^{-n}I_{[0,\theta]}(\max x_i)\cdot\alpha\theta^{-\alpha-1}I_{[\theta_0,\infty)}(\theta)$$

$$=CI_{[\theta_0',\infty)}(\theta)\cdot\theta^{-\alpha-n-1}$$

这里 C 是与 θ 无关的数,$\theta_0'=\max(\theta_0,x_1,\cdots,x_n)$. 可见,后验分布是参数为 $\theta_0',\alpha+n$ 的 Pareto 分布. 证毕.

以上都是讨论单个实参数的先验分布,以下讨论两个实参数的二维先验分布.

定义 3.2 称随机向量(M,R)服从参数为(μ,τ,α,β)的正态 Γ 分布,若 R 的边缘分布是参数为 α,β 的 Γ 分布,而在 $R=r$ 下 M 的条件分布是 $N\left(\mu,\frac{1}{\tau r}\right)$. 这时$(M,R)$的分布密度为

$$\varphi(m,r;\mu,\tau,\alpha,\beta)=\frac{1}{\Gamma(\alpha)}\beta^{\alpha}\sqrt{\tau}\cdot r^{\alpha-\frac{1}{2}}$$

$$\cdot\exp\left\{-\frac{\tau}{2}r(m-\mu)^2-\beta r\right\}$$

定理 3.7 设 $X_1,\cdots,X_n$ 是来自 $N\left(M,\frac{1}{R}\right)$的样本,$M,R$ 木

知. 设(M,R)的先验分布是参数为(μ,τ,α,β)的正态-Γ分布，则在$X_1=x_1,\cdots,X_n=x_n$下，(M,R)的后验分布是参数为$(\mu^*,\tau^*,\alpha^*,\beta^*)$的正态-$\Gamma$分布，这里

$$\mu^*=\frac{\tau\mu+n\bar{x}}{\tau+n},\quad \tau^*=\tau+n$$

$$\alpha^*=\alpha+\frac{n}{2}$$

$$\beta^*=\beta+\frac{1}{2}\sum_1^n(x_i-\bar{x})^2+\frac{\tau n(\bar{x}-\mu)^2}{2(n-1)}$$

$$\left(\bar{x}=\frac{1}{n}\sum_1^n x_i\right)$$

证明：$X_1,\cdots,X_n$的联合密度为

$$f_n(x_1,\cdots,x_n;m,r)=\left(\sqrt{\frac{r}{2\pi}}\right)^n\exp\left\{-\frac{r}{2}\sum_1^n(x_i-m)^2\right\},$$

(M,R)的先验密度为

$$\frac{1}{\Gamma(\alpha)}\beta^\alpha\sqrt{\tau}\,r^{\alpha-\frac{1}{2}}\exp\left\{-\beta r-\frac{\tau r}{2}(m-\mu)^2\right\}$$

从公式(2.3)知(M,R)的后验分布密度为

$$\xi(m,r|x_1,\cdots,x_n)$$
$$=C_1\cdot r^{\alpha+\frac{n}{2}-\frac{1}{2}}\cdot\exp\left\{-\frac{r}{2}\left[\sum_1^n(x_i-m)^2+\tau(m-\mu)^2\right]-\beta r\right\}$$

由于

$$n(\bar{x}-m)^2+\tau(m-\mu)^2$$
$$=(\tau+n)\left(m-\frac{n\bar{x}+\tau\mu}{\tau+n}\right)^2+\frac{n\tau(\bar{x}-\mu)^2}{\tau+n}$$

于是

$$\xi(m,r|x_1,\cdots,x_n)$$
$$=C_1\cdot r^{\alpha+\frac{n}{2}-\frac{1}{2}}\exp\left\{-\frac{\tau+n}{2}r(m-\mu^*)^2-\beta r\right\}$$
$$=\varphi(m,r;\mu^*,\tau^*,\alpha^*,\beta^*)$$

即后验分布是参数为 $\mu^*,\tau^*,\alpha^*,\beta^*$ 的正态-Γ 分布. 证毕.

对于两参数的均匀分布,可以证明它的共轭分布是所谓二维 Pareto 分布.

定义 3.3 称 (Y_1,Y_2) 服从参数是 r_1,r_2,α 的二维 Pareto 分布($r_1<r_2$ 且 $\alpha>0$),若它的分布密度是

$$g(y_1,y_2;r_1r_2,\alpha)=\frac{\alpha(\alpha+1)(r_2-r_1)^\alpha}{(y_2-y_1)^{\alpha+2}}I_{(-\infty,r_1]}(y_1)I_{[r_2,\infty)}(y_2) \tag{3.3}$$

定理 3.8 设 $X_1,\cdots,X_n$ 是来自 $[\theta_1,\theta_2]$ 上均匀分布的样本,$\theta_1<\theta_2$ 未知. 设 (θ_1,θ_2) 的先验分布是参数为 r_1,r_2,α 的二维Pareto 分布,则在 $X=x_1,\cdots,X_n=x_n$ 下,(θ_1,θ_2) 的后验分布是参数为 r_1^*,r_2^*,α^* 的二维 Pareto 分布,这里 $r_1^*=\min(r_1,x_1,\cdots,x_n)$,$r_2=\max(r_2,x_1,\cdots,x_n)$,$\alpha^*=\alpha+n$.

证明:$X_1,\cdots,X_n$ 的联合密度为

$$f_n(x_1,\cdots,x_n;\theta_1,\theta_2)=\prod_1^n\left[\frac{1}{\theta_2-\theta_1}I_{[\theta_1,\theta_2]}(x_i)\right]$$

$$=\frac{1}{(\theta_2-\theta_1)^n}I_{[\theta_1,\infty)}(\min x_i)\cdot I_{(-\infty,\theta_2]}(\max x_i)$$

(θ_1,θ_2) 的先验密度是(3.3). 根据公式(2.3)知 (θ_1,θ_2) 的后验分布密度为

$$\begin{aligned}
&\xi(\theta_1,\theta_2\,|\,x_1,\cdots,x_n)\\
&=\frac{\alpha(\alpha+1)(r_2-r_1)^\alpha}{(\theta_2-\theta_1)^{\alpha+n+2}}I_{(-\infty,r_1]}(\theta_1)I_{[r_2,\infty)}(\theta_2)I_{[\theta_1,\infty)}(\min x_i)\\
&\quad\times I_{(-\infty,\theta_2]}(\max x_i)\\
&=A\,\frac{1}{(\theta_2-\theta_1)^{\alpha+n+2}}I_{(-\infty,r_1^*]}(\theta_1)I_{[r_2^*,\infty)}(\theta_2)\\
&=g(\theta_1,\theta_2;r_1^*,r_2^*,\alpha^*)
\end{aligned}$$

这里 r_1^*,r_2^*,α^* 分别见本定理的叙述. 可见后验分布是参数为 r_1^*,r_2^*,α^* 的二维 Pareto 分布. 证毕.

对于多维分布的共轭分布,我们不做介绍. 读者可参看[11].

对于一般的分布密度(或概率函数)$f(x,\theta)$,如何找其共轭分布族呢？我们指出,只要样本 $X_1,\cdots,X_n$ 有固定维数的充分统计量

$$T(X_1,\cdots,X_n)=(\varphi_1(X_1,\cdots,X_n),\cdots,\varphi_k(X_1,\cdots,X_n))$$

(k 与 n 无关),则有一般方法找出共轭分布族.

设 $X_1,\cdots,X_n$ 的联合密度有下列分解式:

$$\begin{aligned} & f_n(x_1,\cdots,x_n;\theta) \\ & =h(x_1,\cdots,x_n)V_n(T(x_1,\cdots,x_n),\theta) \end{aligned} \tag{3.4}$$

假定

$$0<\int_{\Theta}V_n(t,\theta)\mathrm{d}\theta<\infty ①$$

令

$$\begin{aligned} D=\{(t,n):&存在\ n\geqslant 1\ 及\ x_1,\cdots,x_n\ 使得 \\ & t=(\varphi_1(x_1,\cdots,x_n),\cdots,\varphi_k(x_1,\cdots,x_n))\} \end{aligned}$$

$$g(\theta;t,n)=\frac{V_n(t,\theta)}{\int_{\Theta}V_n(t,\theta)\mathrm{d}\theta}$$

我们可以证明 $G=\{g(\theta;t,n):(t,n)\in D\}$ 是 $f(x,\theta)$ 的共轭分布族.

设 $g(\theta;t,m)\in G$,它是 θ 的先验分布密度,在 $X_1=x_1,\cdots,X_n=x_n$ 下 θ 的后验分布密度是

$$\xi(\theta|x_1,\cdots,x_n)=\frac{f_n(x_1,\cdots,x_n;\theta)g(\theta;t,m)}{\int_{\Theta}f_n(x_1,\cdots,x_n;\theta)g(\theta;t,m)\mathrm{d}\theta} \tag{3.5}$$

设 $X_1,\cdots,X_n,Y_1,\cdots,Y_m$ 独立同分布,则

$$\begin{aligned} & f_{n+m}(x_1,\cdots,x_n,y_1,\cdots,y_m;\theta) \\ & =f_n(x_1,\cdots,x_n;\theta)\cdot f_m(y_1,\cdots,y_m;\theta) \end{aligned}$$

取 $y_1,\cdots,y_m$ 使得 $t=T(y_1,\cdots,y_m)$,对给定的 $x_1,\cdots,x_n$,令 $u=T(x_1,\cdots,x_n,y_1,\cdots,y_m)$,我们来证明:

---

① 更一般地,可用任何别的测度 $N(\mathrm{d}\theta)$ 代替这里的 Lebesgue 测度 $\mathrm{d}\theta$;当 Θ 是至多可数集时,这里及下面的积分均应改为求和.

$$\xi(\theta|x_1,\cdots,x_n)=g(\theta;u,n+m)$$

实际上,从(3.5)知

$$\begin{aligned}
&\xi(\theta|x_1,\cdots,x_n)\\
&=\frac{f_n(x_1,\cdots,x_n;\theta)V_m(t,\theta)h(y_1,\cdots,y_m)}{\int_\Theta f_n(x_1,\cdots,x_n;\theta)\cdot V_m(t,\theta)h(y_1,\cdots,y_m)\mathrm{d}\theta}\\
&=\frac{f_n(x_1,\cdots,x_n;\theta)f_m(y_1,\cdots,y_m;\theta)}{\int_\Theta f_n(x_1,\cdots,x_n;\theta)f_m(y_1,\cdots y_m;\theta)\mathrm{d}\theta}\\
&=\frac{f_{n+m}(x_1,\cdots,x_n,y_1\cdots,y_m;\theta)}{\int_\Theta f_{n+m}(x_1,\cdots,x_n,y_1,\cdots y_m;\theta)\mathrm{d}\theta}\\
&=\frac{V_{n+m}(u,\theta)}{\int_\Theta V_{n+m}(u,\theta)\mathrm{d}\theta}=g(\theta;u,n+m)
\end{aligned}$$

作为本节的末尾,我们介绍广义先验分布的概念:Θ 上任何非负函数 $g(\theta)$ 称为 θ 的广义先验分布密度(又叫权函数). 值得注意的是,虽然 $g(\theta)$ 可能不是概率密度函数,但是公式

$$\xi(\theta|x_1,\cdots,x_n)\stackrel{\mathrm{d}}{=}\frac{f_n(x_1,\cdots,x_n;\theta)g(\theta)}{\int_\Theta f_n(x_1,\cdots,x_n;\theta)g(\theta)\mathrm{d}\theta}$$

(当分母有限时)仍然确定一分布密度,它仍叫做 θ 的后验分布密度.

在实际使用贝叶斯方法时,总要充分利用以往积累的知识,即确定先验分布. 若以往一点知识也没有,怎么办? 此时应如何确定先验分布?

这个重要问题迄今尚未解决,此时没有统一的、公认的方法确定先验分布. 有些人愿意采用下列假设:

(1) 若 θ 在 $(-\infty,\infty)$ 上取值,且无先验知识,则广义先验分布(权函数)应取 $g(\theta)\equiv C(C>0)$.

(2) 若 θ 在 Θ 上取值,$\varphi=\varphi(\theta)$ 是 Θ 到 Ψ 上的一一映射,则 θ 无先验知识时 φ 也无先验知识.

(3) θ 在 (a,b) 上取值 $(-\infty\leqslant a<b\leqslant\infty)$,权函数是 $g(\theta)$,φ 是

(a,b)上有定义的函数，$\varphi'(x)>0$（一切 x）或者 $\varphi'(x)<0$（一切 x），则 $\psi=\varphi(\theta)$的权函数为[1]

$$h(\psi)=g(\varphi^{-1}(\psi))\frac{1}{|\varphi'(\varphi^{-1}(\psi))|} \tag{3.6}$$

这里 φ^{-1}是 φ 的反函数.

根据这三条假设可以推出：①$(0,\infty)$上对应无先验知识的权函数应取$\frac{C}{\theta}$；②$(0,1)$上对应无先验知识的权函数应取$\frac{C}{\theta(1-\theta)}$.

应该指出，上面提到的假设中的第一条和第三条只是一种来自直觉的设想，并无足够的科学依据，从这些假设所导出的结论是否合乎实际还需要通过实践来检验。

*§4 应用实例——电视机寿命验证试验的贝叶斯方法

电视机的使用寿命是很重要的质量指标，所谓使用寿命是指从开始使用到规定的功能丧失为止所经历的时间，当然希望寿命越长越好. 怎样验证一批电视机的寿命是否达到合格标准呢？当然只能采取抽样检查的方法. 这实质上是个假设检验问题，数学上常描述如下. 根据国内外的大量数据，可以认为电视机的寿命 X 服从指数分布，其密度函数如下

$$f(x,\theta)=\begin{cases}0, & x\leqslant 0\\ \frac{1}{\theta}e^{-\frac{1}{\theta}x}, & x>0\end{cases}$$

其中 $\theta\in(0,\infty)$是未知参数. 因为 $EX=\theta$，θ 叫做平均无故障工作时间（记作 MTBF），即平均寿命. 通常取定两个值 $\theta_1,\theta_0,\theta_1<\theta_0$，这个 θ_1 叫检验的下限值（在验收问题中，平均寿命低于 θ_1 时不应接受，θ_1 又叫做极限寿命水平）；θ_0 叫做检验的上限值（平均寿命

---

[1] 当 $g(\theta)$是 θ 的普通概率密度时，从初等概率论知：$\psi=\varphi(\theta)$的概率密度正好由(3.6)给出. 故 $g(\theta)$是权函数时，人们想像 ψ 的权函数也应由(3.6)给出.

高于 θ_0 时应该接受，θ_0 又叫合格寿命水平），比值 $d=\theta_0/\theta_1$ 叫做鉴别比. 传统的试验方案是这样的：从一大批电视机中随机抽取 n 台，同时进行寿命试验，试验 t_0 小时. 在这 $n\times t_0$ 个台时的试验中，电视机的失效数为 r. 当 $r\leqslant C$ 时判定整批合格，使用方接收；当 $r>C$ 时判定整批不合格，使用方拒收. 这个 C 叫做合格判定数，是一个非负整数. 通常称 $T=n\times t_0$ 为总台时数. 这个验证试验方案由三个参数 n,t_0,C 组成. 参数的选择依赖于两类错误带来的风险的允许值. 用 $L(\theta)=P_\theta(r\leqslant C)$ 表示平均寿命是 θ 时接收整批的概率，$L(\theta)$ 叫做抽样特性函数. $\alpha=\sup\limits_{\theta\geqslant\theta_0} P_\theta(r>C)$ 叫做生产方风险，即产品寿命合格却被拒收的概率；$\beta=\sup\limits_{\theta\leqslant\theta_1} P_\theta(r\leqslant C)$ 叫做使用方风险，即产品不合格却被接收的概率. 可以证明 $\alpha=P_{\theta_0}(r>C)$，$\beta=P_{\theta_1}(r\leqslant C)$. 有下列公式：

$$\alpha=1-\sum_{i=0}^{C}\frac{1}{i!}\left(\frac{T}{\theta_0}\right)^i \mathrm{e}^{-\frac{T}{\theta_0}}$$

$$\beta=\sum_{i=0}^{C}\frac{1}{i!}\left(\frac{T}{\theta_1}\right)^i \mathrm{e}^{-\frac{T}{\theta_1}}$$

（理由见(4.2)）. 利用这两个公式就可对给定的 α,β 确定出 T 和 C.

在以往通行的彩色电视机试验中，取 $\theta_1=15\ 000$(h)，$\theta_0=45\ 000$(h)，$n=100$，试验时间 $t_0=1\ 000$(h)，在 $T=n\times t_0=100\ 000$ 台时试验中失效数 $r\leqslant 3$ 时判为合格，这时 $\alpha=0.20$，$\beta=0.10$. 可以看出，在多批交付验收过程中，仍有 20% 的质量合格批被判为质量不合格而遭拒收.

这 100 000 台时的试验实行起来很不容易，费时费电，随着彩电质量的提高，若要验证 MTBF 下限 θ_1 为 20 000 h，则要进行 134 000台时的试验（鉴别比仍为 3，下同）. 若要验证 θ_1 为 30 000 h，则需要 201 000 台时的试验. 上百台彩电同时进行试验需要数百立方米的高温房，仅一次试验就耗电 26 000 度. 可见，这种既费时又费钱的试验方法，显然不适应彩电生产发展的形势.

由上述可见，对于高可靠性产品，用古典统计方法建立的抽样验收试验是不能令人满意的. 我们必须利用已往积累起来的关于平均寿命 θ 的信息. 换句话说，应该另辟捷径，从贝叶斯方法找出路. 我国部分统计工作者于 20 世纪 80 年代中期采用贝叶斯方法对国产彩电制定出新的验证方案，形成了“国家标准”，下面从数学角度对此进行介绍.

(一) 首先找出在固定总台时 T 的试验中失效数 r 服从的概率分布

我们考虑有替换的试验. 设开始时有 n 台电视机投入试验，一旦出现失效，就换一台新的继续试验，试验进行到总台时数为 T 时停止. 设 r 是停止时的累计失效数，则有下列基本定理①.

定理 4.1

$$P_\theta(r=k)=\frac{1}{k!}\left(\frac{T}{\theta}\right)^k e^{-\frac{T}{\theta}} \quad (k=0,1,\cdots) \tag{4.3}$$

这里 P_θ 表示平均寿命是 θ 时相应的概率.

证明：我们首先指出，若 $\xi_1,\xi_2,\cdots,\xi_k$ 相互独立，都服从均值是 θ 的指数分布，则对一切 $t>0$，有

---

① 若试验是无替换的，则可以证明下列公式：

$$P_\theta(r=k)=\frac{1}{k!}\left(\frac{T}{\theta}\right)e^{-\frac{T}{\theta}},k=0,1,2,\cdots,n-1 \tag{4.1}$$

$$P_\theta(r=n)=1-\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{T}{\theta}\right)^k e^{-\frac{T}{\theta}} \tag{4.2}$$

实际上，设 n 个产品的寿命分别为 $X_1,\cdots,X_n$，从小到大排列为 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$. 若 n 个产品同时投入试验，则在时刻 $X_{(i)}$，累计的试验时间为 $S_i=\sum_{k=1}^{i}X_{(k)}+(n-i)X_{(i)}$. 于是对 $k\leqslant n-1$ 有 $P_\theta(r=k)=P_\theta(S_k\leqslant T<S_{k+1})$. 令 $u_k=X_{(k)}-X_{(k-i)}$ ($X_{(0)}\overset{\mathrm{d}}{=}0$)，可以证明 $u_1,u_2,\cdots,u_n$ 相互独立，且 $V_k\overset{\mathrm{d}}{=}(n-k+1)u_k$ 服从均值是 θ 的指数分布

$$P(r=k)=P_\theta\left(\sum_{j=1}^{k}V_j\leqslant T<\sum_{i=1}^{k+1}V_i\right)$$

利用下面的(4.5)即知(4.1)成立，(4.2)从(4.1)直接得到.

$$P\left(\sum_1^k \xi_i \leqslant t\right)=\int_0^t \frac{1}{\theta^k}\mathrm{e}^{-\frac{u}{\theta}} \cdot \frac{u^{k-1}}{(k-1)!}\mathrm{d}u \tag{4.4}$$

可用归纳法证明．当 $k=1$ 时，显然成立．设 $k=l$ 时(4.4)成立，则 $\sum_1^{l+1} \xi_i$ 的密度函数

$$\begin{aligned} p(x) &= \int_0^x \frac{1}{\theta^l}\mathrm{e}^{-\frac{u}{\theta}} \frac{u^{l-1}}{(l-1)!} \cdot \frac{1}{\theta}\mathrm{e}^{-\frac{1}{\theta}(x-u)}\mathrm{d}u \\ &= \frac{1}{\theta^{l+1}}\mathrm{e}^{-\frac{x}{\theta}} \cdot \frac{1}{l!}x^l \end{aligned}$$

可见(4.4)对 $k=l+1$ 也成立．所以对一切 $k\geqslant 1$，(4.4)总成立．

从(4.4)推知

$$\begin{aligned} &P\left(\sum_1^k \xi_i \leqslant t < \sum_1^{k+1} \xi_i\right) \\ &= \frac{1}{k!}\left(\frac{t}{\theta}\right)^k \mathrm{e}^{-\frac{t}{\theta}} \end{aligned} \tag{4.5}$$

设有 n 个产品同时投入试验，若 $\{\xi_{ij}: i=1,2,\cdots,n; j=1,2,\cdots\}$ 是一族相互独立同分布的随机变量，共同分布是均值为 θ 的指数分布．因为试验为有替换的，总试验时间为 T，故试验持续时间为 $t_0=\frac{T}{n}$．显然

$$\begin{aligned} P(r=k) &= \sum_{r_1+\cdots+r_n=k}\prod_{i=1}^n P_\theta\left(\sum_{j=1}^{r_i}\xi_{ij} \leqslant t_0 < \sum_{j=1}^{r_i+1}\xi_{ij}\right) \\ &= \sum_{r_1+\cdots+r_n=k}\prod_{i=1}^n \left[\frac{1}{r_i!}\left(\frac{t_0}{\theta}\right)^{r_i}\mathrm{e}^{-\frac{t_0}{\theta}}\right] \\ &= \frac{1}{k!}\mathrm{e}^{-\frac{T}{\theta}}\sum_{r_1+\cdots+r_n=k}\frac{k!}{r_1!\cdots r_n!}\left(\frac{t_0}{\theta}\right)^{r_1}\cdots\left(\frac{t_0}{\theta}\right)^{r_n} \\ &= \frac{1}{k!}\mathrm{e}^{-\frac{T}{\theta}}\left(\sum_{i=1}^n\left(\frac{t_0}{\theta}\right)\right)^k = \frac{1}{k!}\mathrm{e}^{-\frac{T}{\theta}} \cdot \left(\frac{T}{\theta}\right)^k \end{aligned}$$

这就证明了(4.3)． 证毕．

从(4.3)看出，失效数 r 服从 Poisson 分布. 从(4.1)和(4.2)知，对无替换的试验，失效数近似服从 Poisson 分布.

(二) 先验分布的确定

这是采用贝叶斯方法最关键的一步. 从定理 3.2 知下列逆 Γ 分布①：

$$\xi(\theta)=\frac{b^a}{\Gamma(a)}\theta^{-a-1}\mathrm{e}^{-\frac{b}{\theta}}\quad(\theta>0)\tag{4.6}$$

是 Poisson 分布(4.3)的共轭分布，其中 a,b 是正数.

实际上，若取(4.6)为 θ 的先验分布密度，则在失效数 $r=k$ 的条件下，θ 的后验分布密度为

$$\xi(\theta|k)=\frac{(b+T)^{a+k}}{\Gamma(a+k)}\theta^{-(a+k+1)}\mathrm{e}^{-\frac{b+T}{\theta}}\tag{4.7}$$

故取(4.6)为 θ 的先验分布. 怎样确定 a,b 呢? a,b 的直观意义不明显. 但先验分布的均值 $E\tilde{\theta}$ 及 10%分位数 θ_L(即有 $P(\tilde{\theta}\leqslant\theta_L)=0.10$)有明确的直观意义. 有下列关系式：

$$E\tilde{\theta}=\frac{b}{a-1}\tag{4.8}$$

$$\int_0^{\theta_L}\frac{b^a}{\Gamma(a)}\theta^{-(a+1)}\mathrm{e}^{-\frac{b}{\theta}}\mathrm{d}\theta=0.1\tag{4.9}$$

若能根据历史数据近似估计出 $E\tilde{\theta}$ 及 θ_L，则从(4.8)和(4.9)就可求出 a,b.

为了利用历史数据，对全国 15 个企业生产的 22 种型号彩电的寿命数据进行了调查(包括实验室数据和现场使用数据)，经过多次研究，将产品按寿命分成三级，分别记为 P、Q、R；然后根据管理经验和历史数据确定每级产品的 θ_1,θ_0 的取值，鉴别比 $d=3$. P,Q,R 三级的 θ_1 分别定为 15 000，20 000 和 30 000 h，根据历史数据，取 $E\tilde{\theta}=3\theta_1$. P,Q,R 三级的 θ_L 分别为 $0.75\theta_1$，$0.90\theta_1$，

① 称 η 服从逆 Γ 分布，若 η^{-1} 服从 Γ 分布.

0.95θ_1. 以 P 级为例，取 $E\tilde{\theta}=3\theta_1=45\ 000$；$\theta_L=0.75\theta_1=11\ 250$，再利用(4.8)、(4.9)经计算机算得先验分布中的两个参数如下：

$$a=1.956,\ b=2.868$$

对 Q 级，R 级可进行同样的计算，总之得各级所需的参数如下表所示：

参数 \ 级别	MTBF(平均寿命)		先验分布参数	
	θ_1	θ_0	a	b
P	15 000	45 000	1.956	2.868
Q	20 000	60 000	2.298	3.895
R	30 000	90 000	2.431	4.294

(三) 验证方案的制定

验证方案由(T,C)两个参数构成. T 是总试验时间(台时数)，C 是合格判定数. 当试验停止时累计失效数 $r\leqslant C$，则判定批合格；当 $r>C$ 时，则判定批不合格. T,C 如何确定？应使后验生产方风险 α^* 及后验使用方风险 β^* 都达到指定的水平，这里

$$\alpha^*=P(\tilde{\theta}\geqslant\theta_0\,|\,r>C)$$

$$\beta^*=P(\tilde{\theta}\leqslant\theta_1\,|\,r\leqslant C)$$

可以证明有下列计算公式：

定理 4.2

$$\alpha^*=\frac{\Gamma\left(a,\dfrac{b}{\theta_0}\right)-\left(\dfrac{b}{b+T}\right)^a\displaystyle\sum_{k=0}^{C}\frac{1}{k!}\left(\frac{T}{b+T}\right)^k\Gamma\left(a+k,\frac{b+T}{\theta_0}\right)}{\Gamma(a)[1-P(r\leqslant C)]}\tag{4.10}$$

$$\beta^*=\frac{\left(\dfrac{b}{b+T}\right)^a\displaystyle\sum_{k=0}^{C}\frac{1}{k!}\left(\frac{T}{b+T}\right)^k\left[\Gamma(a+k)-\Gamma\left(a+k,\frac{b+T}{\theta_1}\right)\right]}{P(r\leqslant C)}\tag{4.11}$$

其中 $\Gamma(a,x)=\int_0^x t^{a-1}\mathrm{e}^{-t}\mathrm{d}t$ 是不完全 Γ 函数，而

$$P(r \leqslant C)=\int_0^{\infty} P_\theta(r \leqslant C)\xi(\theta)\mathrm{d}\theta$$

$$=\left(\frac{b}{b+T}\right)^a \sum_{k=0}^{C} \frac{\Gamma(a+k)}{\Gamma(a)k!}\left(\frac{T}{b+T}\right)^k$$

是接收概率，这是一个无条件概率，表示 θ 的先验分布为(4.6)时产品被接收的概率.

证明：

$$\alpha^*=\frac{P(\tilde{\theta} \geqslant \theta_0, r>C)}{P(r>C)}$$

$$=\frac{P(\tilde{\theta} \geqslant \theta_0)-P(\tilde{\theta} \geqslant \theta_0, r \leqslant C)}{P(r>C)}$$

由于

$$P(\tilde{\theta} \geqslant \theta_0)=\int_{\theta_0}^{\infty} \xi(\theta)\mathrm{d}\theta=\int_{\theta_0}^{\infty} \frac{b^a}{\Gamma(a)}\theta^{-(a+1)}\mathrm{e}^{-\frac{b}{\theta}}\mathrm{d}\theta$$

$$=\int_0^{\frac{b}{\theta_0}} \frac{1}{\Gamma(a)} t^{a-1}\mathrm{e}^{-t}\mathrm{d}t$$

$$=\frac{1}{\Gamma(a)}\Gamma\left(a, \frac{b}{\theta_0}\right)$$

另一方面

$$P(\tilde{\theta} \geqslant \theta_0, r \leqslant C)=\int_{\theta_0}^{\infty} P_\theta(r \leqslant C)\xi(\theta)\mathrm{d}\theta$$

$$=\int_{\theta_0}^{\infty} \sum_{k=0}^{C} \frac{1}{k!}\left(\frac{T}{\theta}\right)^k \mathrm{e}^{-\frac{T}{\theta}} \cdot \frac{b^a}{\Gamma(a)}\theta^{-(a+1)}\mathrm{e}^{-\frac{b}{\theta}}\mathrm{d}\theta$$

$$=\frac{1}{\Gamma(a)}\sum_{k=0}^{C} \frac{1}{k!} \frac{T^k b^a}{(T+b)^{a+k}}\int_0^{\frac{T+b}{\theta_0}} t^{a+k-1}\mathrm{e}^{-t}\mathrm{d}t$$

$$=\frac{1}{\Gamma(a)}\left(\frac{b}{T+b}\right)^a \sum_{k=0}^{C} \frac{1}{k!}\left(\frac{T}{T+b}\right)^k \Gamma\left(a+k, \frac{T+b}{\theta_0}\right)$$

由此知(4.10)成立．同理知(4.11)成立．　　　　证毕．

对于 P 级产品，$\theta_1=15\ 000$，$a=1.956$，$b=2.868$，取 $\beta^*=$

0.10，由公式(4.11)可算出几组(T,C)的值，再由公式(4.10)计算相应的后验生产方风险α^*，得下表：

P级产品的贝叶斯方案

C(允许失效数)	1	2	3
T(总试验时间)	25 389	36 421	47 748
α^*(后验生产方风险)	0.044	0.019	0.008

同样可以制定Q级和R级(优质品或名牌产品)的验证试验方案.

经过反复讨论和对历史数据的分析，兼顾使用方和生产方的利益，最后在验收标准中对P级产品选用了$C=2$的方案，这时试验总台时数$T=36\ 421\approx37\ 000$，是原来的经典验收方案台时数$T=100\ 000$的37%，这就大大减少了试验工作量. 对Q级、R级产品也选定了验收方案，总结在下表中：

验收试验方案表

级别 \ (T,C)方案	T/台时	C(合格判定数)
P	37 000	2
Q	32 000	2
R	20 000	1

以上通过实例介绍了如何运用贝叶斯方法，读者可联系§2中的内容加深对贝叶斯方法的了解.

在上面的实例中，只是利用先验分布和后验分布去确定电视机的验证试验方案，使得后验生产方风险α^*及后验使用方风险β^*均达到指定的水平$(\beta^*=0.10,\alpha^*\leqslant0.05)$，并没有去追求"贝叶斯决策"——"风险"最小的验收方案. 要获得贝叶斯决策，只知道先验分布还不够，还必须知道损失函数是怎样的. 在制定验收方案时，选取生产方和使用方都同意的损失函数并不是很简单的事. 由于较难确定"客观"的损失函数，人们常常不去追求贝叶斯决策了. 当然，一旦选好损失函数，则应该追求贝叶斯决策. 以上面的电视机

验证试验为例，如果损失函数是 0－1 损失，则不难求出相应的贝叶斯决策．具体介绍如下：

沿用正文中的各种记号．检验问题是：

$$H_0: \theta \geqslant \theta_0 \leftrightarrow H_1: \theta \leqslant \theta_1 \quad (\theta_0 > \theta_1)$$

设行动空间 $A=\{a_1, a_2\}$，a_1 表示“接收”（即接受假设 H_0），a_2 表示“拒收”（即拒绝假设 H_0）．损失函数 $L(\theta,a)$ 如下：

$$L(\theta,a_1)=\begin{cases}0, & \theta>\theta_1\\ 1, & \theta\leqslant\theta_1\end{cases}$$

$$L(\theta,a_2)=\begin{cases}1, & \theta\geqslant\theta_0\\ 0, & \theta<\theta_0\end{cases}$$

仍取(4.6)为 θ 的先验分布密度，其中参数 a,b 的确定方法见正文．

设 T 是试验的总台时数（有替换的试验），γ 是试验截止时的累计失效数．从定理 4.2 的证明过程知

$$P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=k)=\frac{1}{\Gamma(a+k)}\int_0^{\frac{T+b}{\theta_0}} t^{a+k-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

$$P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=k)=1-\frac{1}{\Gamma(a+k)}\int_0^{\frac{T+b}{\theta_1}} t^{a+k-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

$$(k=0,1,2,\cdots)$$

利用微分法可以证明 $P(\tilde{\theta}\geqslant\theta_0 \mid \gamma=k)$ 是 k 的减函数，从而 $P(\tilde{\theta}\leqslant\theta_1 \mid \gamma=k)$ 是 k 的增函数．故对任何非负整数 C 有

$$\inf_{0\leqslant k\leqslant C} P(\tilde{\theta}\geqslant\theta_0 \mid \gamma=k)=P(\tilde{\theta}\geqslant\theta_0 \mid \gamma=C) \tag{4.12}$$

$$\sup_{k>C} P(\tilde{\theta}\geqslant\theta_0 \mid \gamma=k)=P(\tilde{\theta}\geqslant\theta_0 \mid \gamma=C+1) \tag{4.13}$$

$$\sup_{0\leqslant k\leqslant C} P(\tilde{\theta}\leqslant\theta_1 \mid \gamma=k)=P(\tilde{\theta}\leqslant\theta_1 \mid \gamma=C) \tag{4.14}$$

$$\inf_{k>C} P(\tilde{\theta}\leqslant\theta_1 \mid \gamma=k)=P(\tilde{\theta}\leqslant\theta_1 \mid \gamma=C+1) \tag{4.15}$$

我们指出，对任何非负整数 C，必有 $T=T(C)$ 使得

$$P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=C)=P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=C) \tag{4.16}$$

实际上，令

$$\varphi(T)\xlongequal{\mathrm{d}} P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=C)-P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=C)$$

$$= \frac{1}{\Gamma(a+C)} \int_0^{\frac{T+b}{\theta_0}} t^{a+C-1} \mathrm{e}^{-t} \mathrm{d}t$$

$$+ \frac{1}{\Gamma(a+C)} \int_0^{\frac{T+b}{\theta_1}} t^{a+C-1} \mathrm{e}^{-t} \mathrm{d}t - 1$$

显然 $\varphi(T)$是 T 的严格增连续函数,且 $\varphi(0+)=-1, \varphi(\infty)=1$,故有 $T=T(C)$使得(4.16)成立.

给定 C,设试验的总台时数为 $T=T(C)$,γ 是累计失效数. 令

$$\delta^*(\gamma)=\begin{cases} a_1, & \text{当 } \gamma \leqslant C \\ a_2, & \text{当 } \gamma > C \end{cases}$$

我们来证明,δ^* 是贝叶斯决策.

设 $\xi(\theta|k)$是失效数为 k 时 θ 的后验分布密度,具体表达式见(4.7). 显然有

$$P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=k)=\int_{\theta_0}^{\infty} \xi(\theta|k) \mathrm{d}\theta$$

$$P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=k)=\int_0^{\theta_1} \xi(\theta|k) \mathrm{d}\theta$$

从(4.12)~(4.16)知

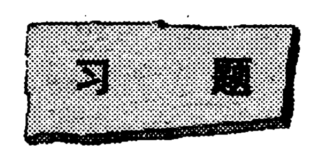
$$\begin{aligned} &\inf_{a\in A}\int_0^{\infty} L(\theta,a)\xi(\theta|k)\mathrm{d}\theta \\ &=\min\{P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=k), P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=k)\} \\ &=\begin{cases} P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=k), & \text{当 } k \leqslant C \\ P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=k); & \text{当 } k > C \end{cases} \end{aligned}$$

另一方面,当 $\gamma=k$ 时 δ^* 的后验平均损失

$$\begin{aligned} \int_0^{\infty} L(\theta,\delta^*(k))\xi(\theta|k)\mathrm{d}\theta &= \begin{cases} P(\tilde{\theta} \leqslant \theta_1 \mid \gamma=k), & k \leqslant C \\ P(\tilde{\theta} \geqslant \theta_0 \mid \gamma=k), & k > C \end{cases} \\ &= \inf_{a\in A}\int_0^{\infty} L(\theta,a)\xi(\theta|k)\mathrm{d}\theta \end{aligned}$$

这表明 δ^* 满足(2.2),故 δ^* 是贝叶斯决策.

在实际工作中要求试验的总台时数尽可能地小,因而 C 应该取得很小. 通常取 $C=0,1,2$,相应的总台时数 $T=T(C)$由(4.16)确定.



习　题

1. 设 X 是离散或连续随机变量,EX^2 存在. 试证明:为了使 $E(X-a)^2$

(a 是实数)达到最小值,必须且只须 $a=EX$.

2. 设 X 是离散或连续随机变量,EX 存在. 试证明:为了使 $E|X-a|$ (a 是实数)达到最小值,必须且只须 a 是 X 的中位数.

3. 设 $X_1,X_2,\cdots,X_n$ 是来自 Poisson 分布的样本,

$$P(X_i=k)=\frac{\lambda^k}{k!}e^{-\lambda}\quad(\lambda>0,k=0,1,\cdots)$$

设 λ 的先验分布是参数为 α,β 的 Γ 分布,损失函数是 $(\lambda-a)^2$,试求 λ 的贝叶斯估计.

4. 设 $X_1,X_2,\cdots,X_n$ 是来自伯努利分布的样本,

$$P(X_i=1)=\theta=1-P(X_i=0),\quad \theta\in(0,1)$$

设 θ 的先验分布是(0,1)上的均匀分布,损失函数是

$$L(\theta,a)=\left[\frac{\theta-a}{\theta(1-\theta)}\right]^2$$

试求 θ 的贝叶斯估计.

5. 设 $X_1,X_2,\cdots,X_n$ 是来自正态分布 $N(\theta,1)$ 的样本,θ 是未知参数. 给定检验问题:

$$H_1:\theta\leqslant\theta_0\leftrightarrow H_2:\theta>\theta_0$$

用 a_i 表示"接受假设 H_i"($i=1,2$). 设损失函数如下:

$$L(\theta,a_1)=\begin{cases}0, & \theta\leqslant\theta_0\\ k(\theta-\theta_0), & \theta>\theta_0\end{cases}$$

$$L(\theta,a_2)=\begin{cases}k|\theta-\theta_0|, & \theta\leqslant\theta_0\\ 0, & \theta>\theta_0\end{cases}$$

其中 k 是正常数. 若 θ 的先验分布是 $N(\mu_0,\sigma_0^2)$,试求出贝叶斯检验.

6. 设 $X_1,X_2,\cdots,X_n$ 是来自 $N(\theta,1)$ 的样本,$\theta\in(-\infty,\infty)$, $A=(-\infty,\infty)$,损失函数 $L(\theta,a)=W(|\theta-a|)(a\in A)$,这里 $W(t)$ 是 $[0,\infty)$ 上的增函数,$W(0)=0$. 若 θ 的先验分布是 $N(0,\sigma^2)$,试求出 θ 的贝叶斯估计. 如果单次观测的费用为 C(正数),问:样本量 n 取多大最好?

7. 地质学家要根据某地区的地层结构来判断该地是否蕴藏石油. 地层结构总是 0,1 两种状态之一;用 θ_0 表示该地无油,θ_1 表示该地有油. 已知有下列概率分布规律(其中 x 表示地层结构的状态,θ 表示石油的状态):

θ \ x	0	1
θ_0（无油）	0.6	0.4
θ_1（有油）	0.3	0.7

它表示如果该地区无油，那么地层结构呈现状态 0 的概率为 0.6，呈现状态 1 的概率为 0.4；如果该地蕴藏石油，那么地层结构呈现状态 0 的概率是 0.3，呈现状态 1 的概率为 0.7. 土地所有者希望根据地质学家对地层结构的分析决定自己投资钻探石油，还是出卖土地所有权或者在该地区开辟旅游点，分别记这些行动为 a_1, a_2, a_3. 行动空间 $A=\{a_1, a_2, a_3\}$，土地所有者权衡利弊之后取损失函数 $L(\theta,a)$ 为

$L(\theta,a)$ \ a / θ	a_1（自己投资钻探）	a_2（出卖所有权）	a_3（开辟旅游点）
θ_0（无油）	12	1	6
θ_1（有油）	0	7	5

试写出可供土地所有者选择的全部决策（函数）及其风险. 求出minimax决策. 若 θ 的先验分布是：$\xi(\theta_0)=0.2, \xi(\theta_1)=0.8$，试求出贝叶斯决策.

第八章 抽样调查概述

§1 问题的提法

抽样调查亦称大规模抽样调查，是数理统计的一个重要内容，其应用极其广泛．无论是政府部门、商业机构、工矿企业、科研单位或其他部门，在制订计划、开发产品、研究现状时都需要许多信息．当现有的数据不充分时，就要进行调查，收集所需的数据．各种各样的调查主要可分为全面调查、典型调查、概率抽样调查三大类．全面调查也称普查，如我国若干年一次的人口普查．这类调查需要大量的人力、物力和时间．典型调查是我国过去用得比较多的调查方法，需要有明确的先验信息，也就是必须事先知道所调查的样本确实是典型的，否则所得资料无太大意义．典型调查一般取样较少，但无法客观地估计所得信息的精度．概率抽样调查是引入概率统计原理的一种调查方法．1935年英国统计学家 Fisher 在选择样本单元时引入了随机化，利用这种方法，不但可获得总体的量的估计，而且对这些估计可估出其可能的错误．今天在概率统计原理指导下，针对实际情况的、合理的概率抽样方案多如繁星．利用它们既可获得可靠的信息，又可大大节约调查的人力、物力、时间．概率抽样调查已成为抽样调查中的主要方法．实践中常将这三类调查配合使用．

传统的概率抽样调查的理论，是考虑从一个由有限个单元组成的总体中，抽取一个样本，根据样本估计总体的一些指标．设有限总体有 N 个单元．记为

$$\mathscr{U}(N)=\{u_1, u_2, \cdots, u_N\}$$

每一单元有对应的指标值，故有 N 个指标值：

$$\{Y_1, Y_2, \cdots, Y_N\}$$

从总体 N 个单元中抽取 n 个单元 $(u_{\theta_1}, u_{\theta_2}, \cdots, u_{\theta_n})$ 作为样本，对样本的各单元测量指标值，得相应的指标值，记为

$$(y_1, y_2, \cdots, y_n)$$

$(y_1, y_2, \cdots, y_n)$ 是 $(Y_1, Y_2, \cdots, Y_N)$ 的一部分. 根据 $(y_1, y_2, \cdots, y_n)$ 估计总体的均值

$$\overline{Y} = \frac{1}{N}(Y_1 + Y_2 + \cdots + Y_N)$$

或总数

$$Y = Y_1 + Y_2 + \cdots + Y_N$$

等数量指标，并研究估计的可能偏差.

这里除了要确定估计量外，另一个重要的问题是如何抽取样本的 n 个单元，也就是一个抽样计划包含设计一个如何取得样本单元的方案以及确定相适应的估计量这两个重要内容.

一个完整的抽样调查除了有一个抽样计划外，也涉及大量的现场调查的实践和组织工作. 整个调查大致有下述内容：

A. 明确调查的目的

B. 调查的准备阶段

1. 总体的确定，明确调查的目标量

2. 明确抽样框(即总体各单元如何与数码 $1, 2, \cdots, N$ 建立对应关系)

3. 收集数据的方法

4. 抽样设计

5. 调查问卷

C. 现场测量阶段

D. 数据处理阶段

1. 数据的验收、编辑

2. 估计、分析

E. 总结报告

一次完美的抽样调查不仅要有一个好的抽样计划，而且要有良好的组织工作、正确的测量手段，每一步骤都应一丝不苟. 否则将会招致调查的失败. 因为抽样调查的误差包含有抽样误差和非抽样误差. 抽样设计主要是应用概率统计原理尽可能地减少抽样误差，而组织工作的疏忽、测量手段的低劣将会使测量失准、数据丢失、记录错误等等非抽样误差大大增加，使最后的估计、分析失去意义.

概率抽样调查中总体一般包含大量的单元，抽取的样本量也是相当大的，因而估计的相合性是一条重要的要求. 对抽样设计中估计量的最基本的准则大致可总结如下：若以 $\tilde{w}$ 估 W.

$$B(\tilde{w})=E(\tilde{w}-W)=E\tilde{w}-W$$

称为偏量.

$$V(\tilde{w})=E(\tilde{w}-W)^2=\mathrm{Var}\,\tilde{w}+[E\tilde{w}-W]^2$$

称为均方偏差或均方误差.

(1) 样本量 n 增大时，偏量与均方偏差应同时变小(若 $N=\infty$，则 $n\to\infty$时，偏量与均方偏差均收敛到 0)，且偏量比均方偏差的平方根变小得更快. 如此，估计量是相合的，且 $\mathrm{Var}\,\tilde{w}$ 可近似作为均方偏差. 这样的估计量称为可用的.

(2) 比较两种估计的好坏，以它们的均方偏差的大小为准，均方偏差小者为佳.

§2 单纯随机抽样

从总体$\{u_1, u_2, \cdots, u_N\}$中抽取样本量为 n 的一个样本. 每次抽取对总体中各单元有相等的被抽中概率，且抽取是不放回的. 即第一次对总体中各单元以 $1/N$ 的概率抽取一个样单元，抽中的单元不放回，对剩下的 $N-1$ 个单元以 $1/(N-1)$的相等概率抽取第二个样单元，……，如此继续，直至抽满 n 个样单元. 这样的抽

样方法通常称为简单随机抽样. 但它与本书前面章节中称为简单随机样本的独立同分布样本不同，它的样本单元间不相互独立. 为了加以区分，以下我们将这种无放回的等概抽样称为单纯随机抽样. 容易证明，采用这种抽样，样本$(y_1, y_2, \cdots, y_n)$的取值范围是所有从$\{Y_1, Y_2, \cdots, Y_N\}$中取 n 个的组合，且取每一组合的概率均相等，为 $1/C_N^n$.

单纯随机抽样是抽样调查中最基本的抽样方法，其他复杂的抽样方法往往以它作为基础. 当我们对被调查的总体知之甚少，没有什么信息可以利用时，往往采用单纯随机抽样.

引进一个与抽样设计有关的函数

$$D_i = \begin{cases} 1, & \text{当单元 } u_i \text{ 进入样本} \\ 0, & \text{当单元 } u_i \text{ 未进入样本} \end{cases}$$

引理 2.1 对单纯随机抽样法，有

$$P\{D_i = 1\} = \frac{n}{N}, \quad i = 1, \cdots, N$$

$$P\{D_i = 1, D_j = 1\} = \frac{n(n-1)}{N(N-1)}, \quad i \neq j,\ i, j = 1, \cdots, N \tag{2.1}$$

证明：在单纯随机抽样下，从 N 个单元中取 n 个单元的一切可能的情形有 C_N^n 种，它们有相等的出现概率. 故由古典概率的算法可得

$$P\{D_i = 1\} = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}, \quad i = 1, \cdots, N$$

$$P\{D_i = 1, D_j = 1\} = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)}, \quad i \neq j$$

证毕.

通过简单的计算，可得

$$E(D_i) = \frac{n}{N}, \quad \mathrm{Var}(D_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$
$$i = 1, \cdots, N$$

$$E(D_i D_j) = \frac{n(n-1)}{N(N-1)},$$

$$\mathrm{Cov}(D_i, D_j) = -\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right)$$

$$i \neq j,\ i,j = 1,\cdots,N \qquad (2.2)$$

上述单纯随机抽样为保证每次抽取对各单元是等概的，可利用随机数表、计算机中的随机数或随机骰子等来实现.

例 2.1(用随机数表实现单纯随机抽样) 调查某一城镇居民用于食物的支出. 该镇有居民共 345 户. 以户为调查单元，拟从 345 户中抽取 20 户作样本. 用随机数表作单纯随机抽样. 首先按户籍册的装订顺序给每户一个对应的编号，号码从 1 至 345. 然后取一随机数表，从中任取三列构成三位的随机数，逐行依次取出不同的三位数. 可规定当数在 001—345 之间，则对应的号码进入样本；当数在 401—745 之间，则该数减去 400 对应的号码进入样本；

随机数表

46	98	63	71	62	33	26	16	80	45	60	11	14	10	95
42	53	32	37	32	27	07	36	07	51	24	51	79	89	73
32	90	79	78	53	13	55	38	58	59	88	97	54	14	10
05	03	72	93	15	57	12	10	14	21	88	26	49	81	76
31	62	43	09	90	06	18	44	32	53	23	83	01	30	30
17	37	93	23	78	87	35	20	96	43	84	26	34	91	64
77	04	74	47	67	21	76	33	50	25	83	92	12	06	76
98	10	50	71	75	12	86	73	58	07	44	39	52	38	79
52	42	07	44	38	15	51	00	13	42	99	66	02	79	54
49	17	46	09	62	90	52	84	77	27	08	02	73	43	28
79	83	86	19	62	06	76	50	03	10	55	23	64	05	05
83	11	46	32	24	20	14	85	88	45	10	93	72	88	71
07	45	32	14	08	32	98	94	07	72	93	85	79	10	75
00	56	76	31	38	80	22	02	53	53	86	60	42	04	53
42	34	07	96	88	54	42	06	87	98	35	85	29	48	39
13	89	51	03	74	17	76	37	13	04	07	74	21	19	30
97	12	25	93	47	70	33	24	03	54	97	77	46	44	80
16	64	36	16	00	04	43	18	66	79	94	77	24	21	90
45	59	34	68	49	12	72	07	34	45	99	27	72	95	14
20	15	37	00	49	52	85	66	60	44	38	68	88	11	80

其余的 000,346—400,746—999 不要;当某号码已进入样本,而再次碰到该号码时,则只算一次. 容易证明当随机数表中号码 000—999 每一号码出现的概率为 1/1 000 时,如此得到的样本是单纯随机样本. 如用下列随机数表取样. 先随意取三列,此处取 8、9、10 三列. 第一个数为 162,则 162 号进入样本. 第二个数为 732,减去 400 为 332,332 号进入样本为第二个样本单元. 第三个数为 853,舍去. 第四个数为 315,则 315 号进入样本为第三个样本单元. 如此继续,可得 175,038,224,008,138,288,200,049 号进入样本. 8、9、10 列随机数已取完,入样的号码仍不足 20,则转入另三列继续取. 如利用 18、19、20 三列,则继续得 045,021,253,243,025,342,327,310,304 号进入样本. 取满 20 个样单元即停止.

另外,实际工作中常使用一种称为系统抽样的技术,按照固定间隔选取样本单元. 如按编号的系统抽样可选取一正整数 k,在号码 $1,2,\cdots,k$ 中,随机地取一个数,若号码 i 进入样本($1\leqslant i\leqslant k$),则 $i,k+i,2k+i,\cdots$ 皆进入样本. 又如调查顾客对某一商品的态度,在商店门口等待顾客,每隔 10 分钟访问一位顾客,这种抽样常常非常简便. 当总体中各单元的编号是完全随机时,系统抽样可看作单纯随机抽样.

定理 2.1 若$(y_1,y_2,\cdots,y_n)$是从$\{Y_1,Y_2,\cdots,Y_N\}$中抽出的单纯随机样本,则

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{2.3}$$

是 $\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$ 的无偏估计. 其均方偏差为

$$\operatorname{Var}(\overline{y})=\frac{1}{n}\left(1-\frac{n}{N}\right)S^2 \tag{2.4}$$

其中 $S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\overline{Y})^2$.

证明：

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{N}Y_iD_i$$

$$E(\bar{y})=\frac{1}{n}\sum_{i=1}^{N}Y_iE(D_i)=\frac{1}{N}\sum_{i=1}^{N}Y_i$$

$$\begin{aligned}
\mathrm{Var}(\bar{y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N}Y_iD_i\right)\\
&=\frac{1}{n^2}\left[\sum_{i=1}^{N}Y_i^2\mathrm{Var}(D_i)+\sum_{i\neq j}\sum Y_iY_j\mathrm{Cov}(D_i,D_j)\right]\\
&=\frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N}Y_i^2\right.\\
&\qquad\left.+\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right)\sum_{i\neq j}\sum Y_iY_j\right]\\
&=\frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\left(1+\frac{1}{N-1}\right)\sum_{i=1}^{N}Y_i^2\right.\\
&\qquad\left.-\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right)\left(\sum_{i=1}^{N}Y_i\right)^2\right]\\
&=\frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}Y_i^2-\frac{1}{N}\left(\sum_{i=1}^{N}Y_i\right)^2\right]\\
&=\frac{1}{n}\left(1-\frac{n}{N}\right)S^2
\end{aligned}$$

证毕.

系 2.1　$N\bar{y}$ 是 $Y=\sum_{i=1}^{N}Y_i$ 的无偏估计. 且

$$\mathrm{Var}(N\bar{y})=\frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2$$

定理 2.2　在单纯随机抽样下

$$v(\bar{y})=\frac{1}{n}\left(1-\frac{n}{N}\right)s^2 \tag{2.5}$$

是 $\mathrm{Var}(\bar{y})$的无偏估计，其中 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

证明：只需证 $$Es^2=S^2.$$

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-n(\bar{y}^2)\right]=\frac{1}{n-1}\left[\sum_{i=1}^{N}Y_i^2D_i-n(\bar{y}^2)\right]$$

$$Es^2=\frac{1}{n-1}\left[\sum_{i=1}^{N}Y_i^2E(D_i)-nE(\bar{y}^2)\right]$$

$$=\frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N}Y_i^2-n\left\{\frac{1}{n}\left(1-\frac{n}{N}\right)S^2+\left(\frac{1}{N}\sum_{i=1}^{N}Y_i\right)^2\right\}\right]$$

$$=\frac{1}{n-1}\left[\frac{n}{N}(N-1)S^2-\left(1-\frac{n}{N}\right)S^2\right]$$

$$=S^2$$

证毕.

定理 2.1 及定理 2.2 告诉我们，在单纯随机抽样下，用

$$\bar{y}\text{估}\bar{Y}\qquad(N\bar{y}\text{估}Y)$$

这种估计的均方偏差 $\mathrm{Var}(\bar{y})=\frac{1}{n}\left(1-\frac{n}{N}\right)S^2$. 这一估计是可用的，且均方偏差有估计量 $v(\bar{y})=\frac{1}{n}\left(1-\frac{n}{N}\right)s^2$.

例 2.2(例 2.1 继续)　对进入样本的 20 户，调查用于食品的月支出，得数据如下(单位：元)：

142	163	142	165	146	178	228
188	169	165	262	163	195	182
160	195	169	167	163	190	

计算得

$$\bar{y}=176.6,\quad s^2=820.57$$

根据上述估计法，估计该镇平均每户的月食品支出为 176.6 元. 这一估计的标准偏差的估计为

$$\sqrt{v(\bar{y})}=\sqrt{\frac{1}{n}\left(1-\frac{n}{N}\right)s^2}=\sqrt{\frac{1}{20}\left(1-\frac{20}{345}\right)(820.57)}$$

$$=6.2(\text{元})$$

§3 区间估计与样本量的确定

利用估计量 $\bar{y}$ 的近似分布，可对 $\bar{Y}$ 做区间估计，其理论基础是下述 Wald-Wolfowitz 定理.

定理 3.1 设 $\{a_{N1},\cdots,a_{NN}\}$ 和 $\{x_{N1},\cdots,x_{NN}\}$，$N=1,2,\cdots$ 是两个实数序列的集合，满足：对 $r=3,4,\cdots$ 及大的 N，有

$$\frac{\frac{1}{N}\sum_{1}^{N}(a_{Ni}-\bar{a}_N)^r}{\left[\frac{1}{N}\sum_{1}^{N}(a_{Ni}-\bar{a}_N)^2\right]^{\frac{r}{2}}}=O(1)$$

$\left(\bar{a}_N \stackrel{\mathrm{d}}{=} \frac{1}{N}\sum_{1}^{N}a_{Ni}\right)$.

$$\frac{\frac{1}{N}\sum_{1}^{N}(x_{Ni}-\bar{x}_N)^r}{\left[\frac{1}{N}\sum_{1}^{N}(x_{Ni}-\bar{x}_N)^2\right]^{\frac{r}{2}}}=O(1)$$

$\left(\bar{x}_N \stackrel{\mathrm{d}}{=} \frac{1}{N}\sum_{1}^{N}x_{Ni}\right)$. 对每一 N，$(X_1,\cdots,X_N)$ 是取值为 $(x_{N1},\cdots,x_{NN})$ 的排列的均匀分布随机向量. 又记

$$L_N=\sum_{1}^{N}a_{Ni}X_i$$

则

$$E(L_N)=N\bar{a}_N\bar{x}_N$$

$$\mathrm{Var}(L_N)=\frac{N^2}{N-1}\left[\frac{1}{N}\sum_{1}^{N}(a_{Ni}-\bar{a}_N)^2\right]\left[\frac{1}{N}\sum_{1}^{N}(x_{Ni}-\bar{x}_N)^2\right]$$

且

$$P\left\{\frac{L_N-E(L_N)}{\sqrt{\mathrm{Var}(L_N)}}\leqslant z\right\}\longrightarrow\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t\qquad(N\to\infty)$$

证明简介：此定理证明中有细致的计算. 细节可参看 A.

Wald and J. Wolfowitz (1944), Statistical tests based on permutations of the observations, Ann. Math. Statist., Vol 15, pp. 358—372. 此处简略地介绍证明步骤. 首先正则化$\{a_{Ni}\}$,$\{x_{Ni}\}$. 令

$$a'_{Ni}=\frac{a_{Ni}-\bar{a}_N}{\left[\frac{1}{N}\sum_1^N(a_{Ni}-\bar{a}_N)^2\right]^{\frac{1}{2}}}$$

$$x'_{Ni}=\frac{x_{Ni}-\bar{x}_N}{\left[\frac{1}{N}\sum_1^N(x_{Ni}-\bar{x}_N)^2\right]^{\frac{1}{2}}}$$

可证$\{a'_{N1},\cdots,a'_{NN}\}$和$\{x'_{N1},\cdots,x'_{NN}\}$满足定理中的类似的条件. 又$(X'_1,\cdots,X'_N)$是取值$(x'_{N1},\cdots,x'_{NN})$的排列的均匀分布随机向量. 令

$$L'_N=\sum_1^N a'_{Ni}X'_i$$

则

$$E(L'_N)=0$$

$$\frac{L'_N}{\sqrt{\mathrm{Var}(L'_N)}}=\frac{L_N-E(L_N)}{\sqrt{\mathrm{Var}(L_N)}}$$

经细致计算可得,对整数 s,

$$E(L'_N)^s=\begin{cases}\frac{(2k)!}{2^k k!}N^k+o(N^k), & 当\ s=2k\\ o(N^k), & 当\ s=2k+1\end{cases}$$

因此

$$E\left(\frac{L'_N}{\sqrt{\mathrm{Var}(L'_N)}}\right)^s=E\left(\frac{L_N-E(L_N)}{\sqrt{\mathrm{Var}(L_N)}}\right)^s$$

$$\longrightarrow\begin{cases}\frac{(2k)!}{2^k k!}, & 当\ s=2k\\ 0, & 当\ s=2k+1\end{cases}\qquad (N\to\infty)$$

即 $N\to\infty$时,$[L_N-E(L_N)]/\sqrt{\mathrm{Var}(L_N)}$的 s 阶矩收敛到标准正态

随机变量的 s 阶矩，因此 $[L_N - E(L_N)]/\sqrt{\mathrm{Var}(L_N)}$，有极限分布 $N(0,1)$.

利用此定理，把 $\{Y_1, \cdots, Y_N\}$ 看成 $\{a_{N1}, \cdots, a_{NN}\}$，$N=1,2,\cdots$ 序列中的一个，而

$$(x_{N1}, \cdots, x_{NN}) = (\underbrace{1, \cdots, 1}_{n}, \underbrace{0, \cdots, 0}_{N-n})$$

且 $\frac{n}{N}=\lambda>0$ 时，$[\bar{y}-\bar{Y}]/\sqrt{\mathrm{Var}(\bar{y})}$ 有近似分布 $N(0,1)$. 又总体的四阶矩 $\frac{1}{N}\sum^{N} Y_i^4$ 有界的话，

$$v(\bar{y})/V(\bar{y}) \xrightarrow{p} 1$$

则 $[\bar{y}-\bar{Y}]/\sqrt{v(\bar{y})}$ 亦近似有分布 $N(0,1)$. 从而给定置信度 $1-\alpha$，可查得标准正态分布的 $1-\frac{\alpha}{2}$ 分位数 $u_{\frac{\alpha}{2}}$，有

$$1-\alpha \doteq P\left\{\left|\frac{\bar{y}-\bar{Y}}{\sqrt{v(\bar{y})}}\right| \leqslant u_{\frac{\alpha}{2}}\right\}$$

$$= P\{\bar{y}-u_{\frac{\alpha}{2}}\sqrt{v(\bar{y})} \leqslant \bar{Y} \leqslant \bar{y}+u_{\frac{\alpha}{2}}\sqrt{v(\bar{y})}\} \qquad (3.1)$$

即 $[\bar{y}-u_{\frac{\alpha}{2}}\sqrt{v(\bar{y})}, \bar{y}+u_{\frac{\alpha}{2}}\sqrt{v(\bar{y})}]$ 是 $\bar{Y}$ 的置信度为 $1-\alpha$ 的区间估计.

例 3.1(例 2.1、2.2 继续)　若要求作出该镇居民每户月平均食品支出的 0.95 置信度的区间估计. 则查出标准正态分布的 $1-\left(\frac{0.05}{2}\right)$ 分位数 $u=1.96$.

$$\bar{y}-u\sqrt{v(\bar{y})} = 176.6-(1.96)(6.2) = 164.4(\text{元})$$

$$\bar{y}+u\sqrt{v(\bar{y})} = 176.6+(1.96)(6.2) = 188.8(\text{元})$$

所求的 0.95 的置信度的置信区间为[164.4,188.8].

在设计抽样调查方案时，取多少个样本单元是一个重要问题. 样本量的大小与抽取样本的方法以及对估计量的精度要求密切相

关．利用正态分布，可粗略确定在单纯随机抽样下为保证估计 $\bar{y}$ 达到一定的精度要求所需的样本量．估计量的精度的提法，通常有绝对精度和相对精度两种．

绝对精度要求：给定精度 d，希望估计量 $\bar{y}$ 与真值 $\bar{Y}$ 有 $|\bar{y}-\bar{Y}|\leqslant d$．$\bar{y}$ 是随机变量，当然只能在一定的置信度 $1-\alpha$ 下保证 $|\bar{y}-\bar{Y}|\leqslant d$，即要求

$$P\{|\bar{y}-\bar{Y}|\leqslant d\}=1-\alpha$$

根据 $[\bar{y}-\bar{Y}]/\sqrt{V(\bar{y})}$ 有近似分布 $N(0,1)$，从而得

$$d\doteq u_{\frac{\alpha}{2}}\sqrt{V(\bar{y})}=u_{\frac{\alpha}{2}}\left[\frac{1}{n}\left(1-\frac{n}{N}\right)S^2\right]^{\frac{1}{2}}$$

解得

$$n\doteq\frac{(u_{\frac{\alpha}{2}})^2S^2}{d^2+\frac{1}{N}(u_{\frac{\alpha}{2}})^2S^2} \tag{3.2}$$

当 N 很大时，可简化为

$$n\doteq\frac{(u_{\frac{\alpha}{2}})^2S^2}{d^2} \tag{3.3}$$

相对精度要求：给定相对精度 h，在置信度 $1-\alpha$ 下，要求

$$P\left\{\left|\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right|\leqslant h\right\}=1-\alpha$$

从而得

$$\bar{Y}h\doteq u_{\frac{\alpha}{2}}\sqrt{V(\bar{y})}=u_{\frac{\alpha}{2}}\left[\frac{1}{n}\left(1-\frac{n}{N}\right)S^2\right]^{\frac{1}{2}}$$

解得

$$n\doteq\frac{(u_{\frac{\alpha}{2}})^2C^2}{h^2+\frac{1}{N}(u_{\frac{\alpha}{2}})^2C^2} \tag{3.4}$$

其中 $C=S/\bar{Y}$ 是总体的变异系数．N 很大时，可简单一点，取

$$n\doteq\frac{(u_{\frac{\alpha}{2}})^2C^2}{h^2} \tag{3.5}$$

以上的式子中的 S^2 和 C^2 实际上是未知数．因而要确定 n，必须有 S^2 或 C^2 的值．实际工作中常采用下列方法获得它们的粗略的估计值．(i)察往法：若总体过去曾被调查过，则采用过去的 S^2 或 C^2 的估计值计算出 n 值．(ii)预查法：先作一个样本量较少的调查．以此小额样本的 S^2 或 C^2 的估计值确定 n 值，然后再追加调查，使总的样本量为 n．

§4 有放回 pps 不等概抽样

在前两节中，我们讨论的单纯随机抽样法对组成总体的各个单元一视同仁，各单元有均等的被抽中的机会，但实际生活中常有总体中各单元不是处于同等地位的情况．例如，例 2.1 中，我们希望调查该镇居民的人均月食品支出，但调查以户为抽样单元，各户的人口数是不同的，人口多的户比人口少的户似乎应受到较大的重视．又如，调查某地区小麦产量，以乡作为抽样单元，但各乡拥有的麦地面积是不一样的．这就要求用不等概抽样，使各单元被抽中的概率正比于所拥有的人口数或麦地面积应是适宜的．这儿的人口数、麦地面积之类的量，我们称为规模测度．

本节介绍一种最简单的不等概抽样——有放回 pps(抽取概率正比于规模测度)抽样．

总体 $\mathscr{U}(N)=\{u_1,\cdots,u_N\}$ 的每一个单元 u_i 除对应有指标量 Y_i 外，还对应有一规模测度 X_i(正整数)，$i=1,\cdots,N$．有放回 pps 抽样的每一次抽取，各单元 u_i 被抽中的概率 P_i 正比于 X_i．每一次抽取后放回被抽中的单元，再进行下一次抽取．具体的实现方法有：

(1) 累积和法

令 $X=\sum\limits_{i=1}^{N}X_i$．对$\{1,2,\cdots,X\}$作有放回简单随机抽样．设抽中的数为 R，则当

$R\in\{1,\cdots,X_1\}$时，u_1 进入样本；

$R\in\{X_1+1,\cdots,X_1+X_2\}$时，$u_2$ 进入样本；

……

$R\in\{(X_1+\cdots+X_{N-1}+1),\cdots,X\}$时，$u_N$ 进入样本.

这一方法原理简单直接，但当 N 较大时，求累积和 X_1,X_1+X_2，$X_1+X_2+X_3,\cdots,X_1+\cdots+X_{N-1},X$ 很麻烦.

(2) 最大规模测度法

令 $M=\max\{X_1,\cdots,X_N\}$. 每一次抽取从 $1,\cdots,N$ 中随机取一数 a；再独立地从 $1,\cdots,M$ 中随机取一数 b. 当 $b\leqslant X_a$ 时，则 u_a 被抽中进入样本. 若 $b>X_a$，则此次无单元被抽中. 如此重复，取满 n 个样本单元为止. 这种抽取法，易见在一次抽取中，各单元 u_a 被抽中进入样本的概率与其规模测度 X_a 成正比. 这种方法省去了计算累积和的麻烦，但取随机数的次数增加了.

例 4.1 一村庄有 8 个果园，分别有果树 50，30，25，40，26，44，20，35 棵. 今以正比于果树棵数的概率，取 3 个果园作样本.

果园编号	1	2	3	4	5	6	7	8
规模测度	50	30	25	40	26	44	20	35

用最大规模测度法取样. $M=50$. 取得随机数对 $(a,b)=(5,17)$，$1\leqslant a\leqslant 8$，$1\leqslant b\leqslant 50$. 因 $17\leqslant 26$，故第 5 号果园进入样本，为第一个样本单元. 再取第二个随机数对，得 $(a,b)=(3,37)$，$1\leqslant a\leqslant 8$，$1\leqslant b\leqslant 50$. 因 $37>25$，故此次无单元进入样本. 再取随机数对 $(a,b)=(6,18)$，$1\leqslant a\leqslant 8$，$1\leqslant b\leqslant 50$. 因 $18\leqslant 44$，故第 6 号果园进入样本，为第二个样本单元. 再取随机数对 $(a,b)=(2,23)$，$1\leqslant a\leqslant 8$，$1\leqslant b\leqslant 50$. 因 $23\leqslant 30$，故第 2 号果园进入样本，为第三个样本单元. 样本量取满，抽样结束.

定理 4.1 在有放回 pps 抽样下，若样本记为

$$(y_1,y_2,\cdots,y_n)$$

样本单元 y_i 是总体 $\mathscr{U}(N)$ 的某一单元，其对应的抽取概率记为 $p_i, i=1,\cdots,n$. 则

$$\hat{Y}_{\text{pps}}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \tag{4.1}$$

是总体总数 $Y=\sum_{i=1}^{N}Y_i$ 的无偏估计. 其方差为

$$V(\hat{Y}_{\text{pps}})=\frac{1}{n}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-Y\right)^2 \tag{4.2}$$

其中 P_i 为 Y_i 所对应的抽取概率.

证明:令 $\qquad z_i=y_i/p_i, \quad i=1,\cdots,n$

则 $(z_1,\cdots,z_n)$ 独立同分布

$$P\{Z=Y_i/P_i\}=P_i, \quad i=1,\cdots,N$$

故

$$E(\hat{Y}_{\text{pps}})=E\left(\frac{1}{n}\sum_{i=1}^{n}z_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(z_i)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{i=1}^{N}\frac{Y_i}{P_i}\cdot P_i\right)=Y$$

$$\text{Var}(\hat{Y}_{\text{pps}})=\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}z_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(z_i)$$

$$=\frac{1}{n}\text{Var}(z_1)=\frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}-Y\right)^2\cdot P_i$$

证毕.

定理 4.2 在有放回 pps 抽样下，$V(\hat{Y}_{\text{pps}})$ 的一个无偏估计量为

$$v(\hat{Y}_{\text{pps}})=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\hat{Y}_{\text{pps}}\right)^2 \tag{4.3}$$

证明:考虑 $z_1,\cdots,z_n$ 是独立同分布样本. 样本方差

$$\frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\hat{Y}_{\text{pps}}\right)^2$$

是分布

$$P\{Z=Y_i/P_i\}=P_i,\quad i=1,\cdots,N$$

的方差

$$\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}-Y\right)^2\cdot P_i$$

的无偏估计. 故有本定理之结论. 证毕.

例 4.2(例 2.1 继续) 拟调查该镇居民人均月食品支出. 该镇居民总数为 1 300 人. 各户的人数可从户籍册获得. 以正比于各户人数的抽取概率从 345 户中作有放回 pps 抽样,抽取 20 户作样本. 得数据如下:

样单元号	人数	支出	样单元号	人数	支出	样单元号	人数	支出
1	2	142	8	2	188	15	4	160
2	4	163	9	4	169	16	4	195
3	3	142	10	3	165	17	4	169
4	2	165	11	5	262	18	4	167
5	2	146	12	3	163	19	2	163
6	5	178	13	6	195	20	4	190
7	4	228	14	3	182			

抽取概率 P_i 正比于各户人数,也就是 $P_i=$(该户人数)/(总人数). 故该镇月食品支出总数的估计值为

$$\begin{aligned}\hat{Y}_{\text{pps}}&=\frac{1}{n}\sum_{i=1}^{n}y_i/p_i\\&=\frac{1}{20}\left[\frac{142}{2/1\,300}+\frac{163}{4/1\,300}+\cdots+\frac{190}{4/1\,300}\right]\\&=71\,505.42(\text{元})\end{aligned}$$

人均月食品支出的估计值为

$$\hat{Y}_{\text{pps}}/N=\frac{71\,505.42}{1\,300}=55.00(\text{元})$$

对应的标准偏差的估计值为

$$\left[\frac{1}{N^2}\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\hat{Y}_{\text{pps}}\right)^2\right]^{\frac{1}{2}}$$

$$=\left[\frac{1}{N^2}\frac{1}{n(n-1)}\left\{\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2-n\hat{Y}_{\text{pps}}^2\right\}\right]^{\frac{1}{2}}$$

$$=\left[\frac{1}{20\cdot 19}\left\{\left(\frac{142}{2}\right)^2+\left(\frac{163}{4}\right)^2+\cdots+\left(\frac{190}{4}\right)^2\right.\right.$$

$$\left.\left.-20\cdot(55.00)^2\right\}\right]^{\frac{1}{2}}$$

$=3.83$(元)

§5 分层抽样

若总体 $\mathscr{U}(N)$ 的单元可集合成若干互不相交的子总体，各子总体中的单元的特性较为相近，从每一子总体中取一些样单元组成总的样本. 这样的样本往往能较好地反映总体的情况，比直接从总体中随机地抽取的样本更有代表性. 例如调查全国成年人的平均身高，总体是全国的成年人. 由于我国幅员辽阔，各地人的体型有一定的差异，因而按人类学的特征，将全国分成西南、南方、长江流域、黄河流域、北方等几个子总体. 在各个子总体内较之子总体之间人的体型差异要小得多. 又如调查民众对某一问题的意见. 将民众按其职业分成国家干部、工人、农民、家庭主妇等等子总体，则子总体内的意见也较为接近. 这样在各子总体中取一样本组成总的样本更能反映总体的特性. 这种在各子总体中取一样本组成总的样本的方法称为分层抽样，子总体称为层.

设总体

$$\mathscr{U}(N)=\mathscr{U}(N_1)\cup\cdots\cup\mathscr{U}(N_k),\quad N_1+\cdots+N_k=N$$

第 i 层(子总体) $\mathscr{U}(N_i)=\{Y_{i1},\cdots,Y_{iN_i}\}$

从每一层中抽取一样本

$$y_{i1},\cdots,y_{in_i},\quad i=1,\cdots,k$$

k 个层的样本合在一起组成总的样本. 样本量 $n=n_1+\cdots+n_k$. 记

$$W_i=\frac{N_i}{N},\quad w_i=\frac{n_i}{n}$$

$$\overline{Y}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}Y_{ij},\quad S_i^2=\frac{1}{N_i-1}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i)^2$$

$$\overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij},\quad s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$$

$$\overline{Y}=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}Y_{ij}$$

定理 5.1 当各层均是单纯随机样本时,

$$\overline{y}_{st}=\sum_{i=1}^{k}W_i\overline{y}_i \tag{5.1}$$

是 $\overline{Y}$ 的无偏估计. 其均方偏差为

$$V(\overline{y}_{st})=\sum_{i=1}^{k}W_i^2\frac{1}{n_i}\left(1-\frac{n_i}{N_i}\right)S_i^2 \tag{5.2}$$

证明:由于

$$\overline{Y}=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}Y_{ij}=\frac{1}{N}\sum_{i=1}^{k}N_i\overline{Y}_i=\sum_{i=1}^{k}W_i\overline{Y}_i$$

故立即可得定理之结论.

证毕.

定理 5.2 当各层均是单纯随机样本时,$V(\overline{y}_{st})$的一个无偏估计量为

$$v(\overline{y}_{st})=\sum_{i=1}^{k}W_i^2\frac{1}{n_i}\left(1-\frac{n_i}{N_i}\right)s_i^2 \tag{5.3}$$

证明:因 s_i^2 是 S_i^2 的无偏估计,故有定理之结论.

证毕.

分层抽样时,样本量的分配是需要考虑的. 即给定总的样本量 n,如何分配给各个层,定出层的样本量 n_i. 下面介绍几种分配方法.

(1) 等额分配. 即每一层取样本量 $n_i=n/k$. 这样分配样本量

主要是为了管理方便. 如每一层由一个调查员负责. 则等额分配样本使各调查员的工作量一致.

(2) 按比例分配. 即取

$$n_i = n \cdot \frac{N_i}{N}, \quad i=1,\cdots,k$$

这种分配法使各层中的单元进入样本的概率相同. 此时估计量

$$\bar{y}_{st} = \sum_{i=1}^{k} W_i \bar{y}_i = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

即 $\bar{y}_{st}$ 是全部样本单元的简单平均. 这样的样本在进行数据分析时常使处理大为简化.

(3) 奈曼最优分配. 考虑 $n = \sum_{i=1}^{k} n_i$ 的分配,使估计量的均方偏差

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{1}{n_i}\left(1 - \frac{n_i}{N_i}\right) S_i^2$$

达到最小. 用数学分析中解条件极值的方法,可求得

$$n_i = n \cdot \frac{W_i S_i}{\sum_{l=1}^{k} W_l S_l}, \quad i = 1,\cdots,k \tag{5.4}$$

实际求出 n_i 必须知道 S_i 间的比例. 一般多采用察往法或预查法获得 S_i 间的近似比例.

分层抽样的目的是提高估计的精度和方便管理. 例如在全国范围抽样时,以每个省为一层. 层与行政区划一致将极有利于调查的组织管理工作. 为了使估计有较高的精度,则应使各层内的单元比较匀一. 这一点可以从下面的例 5.1 中明显地看出来.

例 5.1 下表是 1984 年我国大陆除西藏自治区外 28 个省市自治区的工业产值(亿元):

1	246.07	8	104.38	15	63.93	22	36.54
2	204.26	9	100.86	16	63.10	23	34.27
3	184.95	10	98.30	17	60.17	24	31.39
4	167.43	11	96.66	18	45.62	25	28.41
5	133.36	12	87.59	19	39.22	26	22.35
6	126.70	13	75.84	20	37.77	27	6.23
7	120.71	14	73.52	21	37.18	28	5.95

假定从 28 个省市自治区中抽取 12 个作为样本，来估计全国工业总产值.

（1）若直接从 28 个省市自治区中用单纯随机抽样取 12 个样单元. 则估计为 $\hat{Y}_{SE}=N\,\bar{y}$. 其均方偏差由 § 2 知

$$V(\hat{Y}_{SE})=N^2\,\frac{1}{n}\left(1-\frac{n}{N}\right)S^2$$

从上表的 28 个数据可算出

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\overline{Y})^2=3\,721.00$$

故可得均方偏差

$$V(\hat{Y}_{SE})=(28)^2\,\frac{1}{12}\left(1-\frac{12}{28}\right)(3\,721.00)=138\,917.33$$

（2）若将 28 个省市自治区分为两层，产值较大的前 14 个为第一层，产值较小的后 14 个为第二层. 按比例分配样本量. 从第一层用单纯随机抽样取 6 个样单元. 从第二层用单纯随机抽样取 6 个样单元. 则全国工业总产值的分层估计的均方偏差为

$$V(N\,\bar{y}_{st})=\sum_{i=1}^{k}N^2W_i^2\,\frac{1}{n_i}\left(1-\frac{n_i}{N_i}\right)S_i^2$$

现在 $k=2, W_1=\frac{14}{28}=\frac{1}{2}, W_2=\frac{1}{2}, n_1=n_2=6$，从前面表中的 28 个数据可算出 $S_1^2=2\,696.72, S_2^2=327.61$. 故

$$V(N\bar{y}_{st})=(28)^2\left[\frac{1}{4}\cdot\frac{1}{6}\cdot\left(1-\frac{6}{14}\right)(2\,696.72)\right.$$

$$\left.+\frac{1}{4}\cdot\frac{1}{6}\left(1-\frac{6}{14}\right)(327.61)\right]$$

$$=56\,454.16$$

较(1)中之值小.

(3) 若采用奈曼最优分配. 在(2)中的两个层中作分层抽样,则理想的样本量分配为

$$n_1=n\frac{W_1S_1}{W_1S_1+W_2S_2}=12\cdot\frac{\left(\frac{1}{2}\right)(51.9)}{\left(\frac{1}{2}\right)(51.9)+\left(\frac{1}{2}\right)(18.1)}\doteq 9$$

$$n_2=n\frac{W_2S_2}{W_1S_1+W_2S_2}\doteq 3$$

按上述样本量在两个层中作单纯随机抽样,则全国工业总产值的估计的均方偏差为

$$(28)^2\left[\left(\frac{1}{4}\right)\left(\frac{1}{9}\right)\left(1-\frac{9}{14}\right)(2\,696.72)\right.$$

$$\left.+\left(\frac{1}{4}\right)\left(\frac{1}{3}\right)\left(1-\frac{3}{14}\right)(327.61)\right]$$

$$=37\,791.80$$

§6 二阶抽样

二阶抽样或三阶、四阶抽样在大型的实际调查工作中是常用的方法. 例如要调查某省农户的某些指标,则各农户即为最小单元. 但全省农户数量太大,且各农户的档案归所在乡镇掌管,因而若直接以农户为抽取单元无现成的抽样框可以利用. 这时可采用二阶抽样. 抽样分二"阶段"进行. 首先将全省的乡镇编号作为第一性抽样单元,在全省中选定若干个乡作为第一性样本单元. 然后再在被抽中的乡中将农户编号作为第二性抽样单元,选取若

干农户为第二性样本单元，调查它们的数量指标. 采用二阶抽样使被调查的农户集中于若干个乡，因而较全省范围的随机抽样，有便于管理、省力、省钱等优点. 但由于被调查的农户集中于若干个乡中，因而当同一个乡内之农户间的差异小于乡之间的差异时，二阶样本不如全省范围的随机样本有代表性.

二阶抽样问题的提法如下：总体

$$\mathscr{U}(N)=\mathscr{U}(N_1)\cup\cdots\cup\mathscr{U}(N_k),\quad N_1+\cdots+N_k=N$$

$\mathscr{U}(N_i), i=1,\cdots,k$ 为第一性抽样单元. 它有 N_i 个第二性抽样单元

$$\mathscr{U}(N_i)=\{Y_{i1},\cdots,Y_{iN_i}\},\quad i=1,\cdots,k$$

调查的总目标仍是总体的总和 $Y=\sum_{i=1}^{k}\sum_{j=1}^{N_i}Y_{ij}$ 或总体的平均值 $\overline{Y}=Y/N$ 等等. 它们是各个第一性抽样单元的目标量 $Y_i=\sum_{j=1}^{N_i}Y_{ij}$ 的和或 $\overline{Y}_i$ 的线性组合. 我们可一般地将调查的总目标量记为 G，各第一性抽样单元对应的目标量记为 $G_i, i=1,\cdots,k$. 不妨假定 $G=G_1+\cdots+G_k$. 因为第一性单元在第一阶段抽样中均有可能被抽中，所以对每个第一性单元拟定一个组内抽样计划，并选定目标量 G_i 的估计量. 给以适当的记号列于下表：

第一性单元	$\mathscr{U}(N_1)$,	$\cdots$,	$\mathscr{U}(N_k)$
第一性目标量	G_1	$\cdots$,	G_k
拟定的第二阶段的组内抽样法	方法 1	$\cdots$,	方法 k
G_i 的估计量	g_1,	$\cdots$,	g_k
g_i 的均方偏差	σ_1^2,	$\cdots$,	σ_k^2
σ_i^2 的估计量	$\hat{\sigma}_1^2$,	$\cdots$,	$\hat{\sigma}_k^2$

此外，还需选定第一阶段的抽样法. 常用的有(i)单纯随机抽样，(ii)有放回 pps 抽样. 对每一个抽中的第一性单元，按拟定的组内

抽样法抽取第二性单元做调查，算出 g_i，以它们为基础估计 G，并研究估计量的均方偏差.

定理 6.1 当第一阶段采用单纯随机抽样，从 K 个第一性单元中抽取 k 个. 抽中的号码是 $\theta_1, \cdots, \theta_k$. 且拟定的组内估计量 g_i 满足：$E(g_i)=G_i, E(\hat{\sigma}_i^2)=\sigma_i^2, i=1,\cdots,k$. 则

(1) $$\hat{G}_s=\frac{K}{k}\sum_{i=1}^{k} g_{\theta_i} \tag{6.1}$$

是 G 的无偏估计量.

(2) $\hat{G}_s$ 的均方偏差为

$$V(\hat{G}_s)=\frac{K^2}{k}\left(1-\frac{k}{K}\right)\left[\frac{1}{K-1}\sum_{i=1}^{K}\left(G_i-\frac{G}{K}\right)^2\right]+\frac{K}{k}\sum_{i=1}^{K}\sigma_i^2 \tag{6.2}$$

(3) $V(\hat{G}_s)$的一个无偏估计量为

$$v(\hat{G}_s)=\frac{K^2}{k}\left(1-\frac{k}{K}\right)\left[\frac{1}{k-1}\sum_{i=1}^{k}\left(g_{\theta_i}-\frac{1}{k}\sum_{l=1}^{k} g_{\theta_l}\right)^2\right]+\frac{K}{k}\sum_{i=1}^{k}\hat{\sigma}_{\theta_i}^2 \tag{6.3}$$

证明：抽得之组号 $\theta_1,\cdots,\theta_k$ 是随机变量. 但对任何抽定之组号，$g_{\theta_1},\cdots,g_{\theta_k}$ 是相互独立的估计量. 我们用 $E(t|\theta)$表示在 $\theta_1,\cdots,\theta_k$ 取定的条件下的数学期望. 有

$$E(g_{\theta_i}|\theta)=G_{\theta_i},$$
$$E(g_{\theta_i}^2|\theta)=\sigma_{\theta_i}^2+G_{\theta_i}^2,\quad i=1,\cdots,k$$
$$E(g_{\theta_i}g_{\theta_j}|\theta)=G_{\theta_i}G_{\theta_j},\quad i\neq j,\ i,j=1,\cdots,k$$

从而

(1) $$E(\hat{G}_s)=E[E(\hat{G}_s|\theta)]=E\left[\frac{K}{k}\sum_{i=1}^{k} G_{\theta_i}\right]$$
$$=\frac{K}{k}\sum_{i=1}^{k}E(G_{\theta_i})=\frac{K}{k}\sum_{i=1}^{k}\left[\sum_{j=1}^{K}G_jP\{\theta_i=j\}\right]$$
$$=\frac{K}{k}\sum_{i=1}^{k}\left[\sum_{j=1}^{K}G_j\frac{1}{K}\right]=G$$

(2) $E(\hat{G}_s^2)=E[E(\hat{G}_s^2|\theta)]$

$$=E\left[\left(\frac{K}{k}\right)^2\left\{\sum_{i=1}^{k}E(g_{\theta_i}^2|\theta)+\sum_{i\neq j}\sum E(g_{\theta_i}g_{\theta_j}|\theta)\right\}\right]$$

$$=E\left[\left(\frac{K}{k}\right)^2\left\{\sum_{i=1}^{k}(\sigma_{\theta_i}^2+G_{\theta_i}^2)+\sum_{i\neq j}\sum G_{\theta_i}G_{\theta_j}\right\}\right]$$

$$=\left(\frac{K}{k}\right)^2\left[\sum_{i=1}^{k}E(\sigma_{\theta_i}^2)+\sum_{i=1}^{k}E(G_{\theta_i}^2)\right.$$

$$\left.+\sum_{i\neq j}\sum E(G_{\theta_i}G_{\theta_j})\right]$$

而

$$\sum_{i=1}^{k}E(\sigma_{\theta_i}^2)=\sum_{i=1}^{k}\sum_{j=1}^{K}\sigma_j^2P\{\theta_i=j\}=\frac{k}{K}\sum_{j=1}^{K}\sigma_j^2$$

$$\sum_{i=1}^{k}E(G_{\theta_i}^2)=\sum_{i=1}^{k}\sum_{j=1}^{K}G_j^2P\{\theta_i=j\}=\frac{k}{K}\sum_{j=1}^{K}G_j^2$$

$$\sum_{i\neq j}\sum E(G_{\theta_i}G_{\theta_j})=\sum_{i\neq j}\sum\left[\sum_{l\neq m}^{K}\sum^{K}G_lG_mP\{\theta_i=l,\theta_j=m\}\right]$$

$$=\frac{k(k-1)}{K(K-1)}\sum_{l\neq m}^{K}\sum^{K}G_lG_m$$

故

$$E(\hat{G}_s^2)=\left(\frac{K}{k}\right)^2\left[\frac{k}{K}\sum_{j=1}^{K}\sigma_j^2+\frac{k}{K}\sum_{j=1}^{K}G_j^2+\frac{k(k-1)}{K(K-1)}\sum_{i\neq j}^{K}\sum^{K}G_iG_j\right]$$

$$=\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2+\left(\frac{K}{k}\right)\left[\left(1-\frac{k-1}{K-1}\right)\sum_{j=1}^{K}G_j^2\right.$$

$$\left.+\frac{k-1}{K-1}\left(\sum_{j=1}^{K}G_j\right)^2\right]$$

$$=\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2+\left(\frac{K}{k}\right)\left[\left(1-\frac{k-1}{K-1}\right)\left(\sum_{j=1}^{K}G_j^2-\frac{G^2}{K}\right)\right.$$

$$\left.+\left(\frac{k-1}{K-1}+\frac{1}{K}\left(1-\frac{k-1}{K-1}\right)\right)G^2\right]$$

$$=\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2+\frac{K^2}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2+G^2$$

所以

$$V(\hat{G}_s)=E(\hat{G}_s^2)-G^2$$

$$=\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2+\frac{K}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2$$

（3）有

$$E(g_{\theta_i}^2)=E[E(g_{\theta_i}^2\mid\theta)]=E[\sigma_{\theta_i}^2+G_{\theta_i}^2]$$

$$=\frac{1}{K}\sum_{j=1}^{K}\sigma_j^2+\frac{1}{K}\sum_{j=1}^{K}G_j^2$$

$$E(\hat{\sigma}_{\theta_i}^2)=E[E(\hat{\sigma}_{\theta_i}^2\mid\theta)]=E(\sigma_{\theta_i}^2)=\frac{1}{K}\sum_{j=1}^{K}\sigma_j^2$$

从而

$$E\left[\frac{1}{k-1}\sum_{i=1}^{k}\left(g_{\theta_i}-\frac{1}{k}\sum_{l=1}^{k}g_{\theta_l}\right)^2\right]$$

$$=\frac{1}{k-1}\left[\sum_{i=1}^{k}E(g_{\theta_i}^2)-kE\left(\frac{1}{k}\sum_{l=1}^{k}g_{\theta_l}\right)^2\right]$$

$$=\frac{1}{k-1}\left[\sum_{i=1}^{k}E(g_{\theta_i}^2)-\frac{k}{K^2}E(\hat{G}_s^2)\right]$$

$$=\frac{1}{k-1}\left[\frac{k}{K}\sum_{j=1}^{K}\sigma_j^2+\frac{k}{K}\sum_{j=1}^{K}G_j^2-\frac{k}{K^2}\left\{\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2\right.\right.$$

$$\left.\left.+\frac{K^2}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2+G^2\right\}\right]$$

$$=\frac{1}{K}\sum_{j=1}^{K}\sigma_j^2+\frac{1}{k-1}\left[\left(\frac{k}{K}-\frac{K-k}{K(K-1)}\right)\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2\right]$$

$$=\frac{1}{K}\sum_{j=1}^{K}\sigma_j^2+\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2$$

所以

$$E[v(\hat{G}_s)]=\frac{K}{k}\left(1-\frac{k}{K}\right)\left[\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2\right.$$

$$\left.+\frac{1}{K}\sum_{j=1}^{K}\sigma_j^2\right]+\frac{K}{k}\left[\frac{k}{K}\sum_{j=1}^{K}\sigma_j^2\right]$$

$$=\frac{K}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2$$

$$+\left(1+\frac{K-k}{k}\right)\sum_{j=1}^{K}\sigma_j^2$$

$$=\frac{K}{k}\left(1-\frac{k}{K}\right)\frac{1}{K-1}\sum_{j=1}^{K}\left(G_j-\frac{G}{K}\right)^2+\frac{K}{k}\sum_{j=1}^{K}\sigma_j^2$$

$$=V(\hat{G}_s)$$

证毕.

定理 6.2 当第一阶段采用有放回 pps 抽样，抽中的号码为 $\theta_1,\cdots,\theta_k$，对应的抽取概率分别为 $P_{\theta_1},\cdots,P_{\theta_k}$. 且拟定的组内估计量 g_i 满足：$E(g_i)=G_i$，$E(\hat{\sigma}_i^2)=\sigma_i^2$，$i=1,\cdots,K$. 则

(1) $$\hat{G}_{\text{pps}}=\frac{1}{k}\sum_{i=1}^{k}\frac{g_{\theta_i}}{P_{\theta_i}} \tag{6.4}$$

是 G 的无偏估计量.

(2) $\hat{G}_{\text{pps}}$的均方偏差为

$$V(\hat{G}_{\text{pps}})=\frac{1}{k}\sum_{i=1}^{K}\frac{\sigma_i^2}{P_i}+\frac{1}{k}\sum_{i=1}^{K}P_i\left(\frac{G_i}{P_i}-G\right)^2 \tag{6.5}$$

(3) $V(\hat{G}_{\text{pps}})$的一个无偏估计量为

$$v(\hat{G}_{\text{pps}})=\frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\frac{g_{\theta_i}}{P_{\theta_i}}-\hat{G}_{\text{pps}}\right)^2 \tag{6.6}$$

证明：令 $\qquad Z_i=g_{\theta_i}/P_{\theta_i},\quad i=1,\cdots,k$

由于抽样是有放回 pps 抽样，故 $\theta_1,\cdots,\theta_k$ 相互独立同分布. 从而 $Z_1,\cdots,Z_k$ 独立，且

$$E(Z_i)=E\left(\frac{g_{\theta_i}}{P_{\theta_i}}\right)=E\left[E\left(\frac{g_{\theta_i}}{P_{\theta_i}}\middle|\theta\right)\right]=E\left[\frac{1}{P_{\theta_i}}G_{\theta_i}\right]$$

$$=\sum_{j=1}^{K}\frac{G_j}{P_j}P\{\theta_i=j\}=\sum_{j=1}^{K}\frac{G_j}{P_j}\cdot P_j=\sum_{j=1}^{K}G_j=G$$

$$\begin{aligned}\mathrm{Var}(Z_i)&=E(Z_i^2)-(EZ_i)^2\\&=E\left[E\left\{\left(\frac{g_{\theta_i}}{P_{\theta_i}}\right)^2\Big|\theta\right\}\right]-G^2\\&=E\left[\left(\frac{1}{P_{\theta_i}}\right)^2(\sigma_{\theta_i}^2+G_{\theta_i}^2)\right]-G^2\\&=\sum_{j=1}^{K}\left(\frac{1}{P_j}\right)^2(\sigma_j^2+G_j^2)\cdot P_j-G^2\\&=\sum_{j=1}^{K}\frac{\sigma_j^2}{P_j}+\sum_{j=1}^{K}\frac{G_j^2}{P_j}-G^2\\&=\sum_{j=1}^{K}\frac{\sigma_j^2}{P_j}+\sum_{j=1}^{K}P_j\left(\frac{G_j}{P_j}-G\right)^2\end{aligned}$$

因而有

(1) $E(\hat{G}_{\mathrm{pps}})=\dfrac{1}{k}\displaystyle\sum_{i=1}^{k}E(Z_i)=G$

(2) $$V(\hat{G}_{\mathrm{pps}})=\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k}Z_i\right)=\frac{1}{k^2}\sum_{i=1}^{k}\mathrm{Var}(Z_i)$$

$$=\frac{1}{k}\sum_{j=1}^{K}\frac{\sigma_j^2}{P_j}+\frac{1}{k}\sum_{j=1}^{K}P_j\left(\frac{G_j}{P_j}-G\right)^2$$

(3) 由于 $Z_1,\cdots,Z_k$ 独立同方差，故 $\dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k}(Z_i-\bar{Z})^2$ 是 $\mathrm{Var}(Z_i)$ 的无偏估计，即 $\dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k}\left(\frac{g_{\theta_i}}{P_{\theta_i}}-\hat{G}_{\mathrm{pps}}\right)^2$ 是

$$\mathrm{Var}(Z_i)=\sum_{j=1}^{K}\frac{\sigma_j^2}{P_j}+\sum_{j=1}^{K}P_j\left(\frac{G_j}{P_j}-G\right)^2$$

的无偏估计. 从而有定理之结论(3). 证毕.

例 6.1 前面例 4.1 中，调查该村庄水果总产量. 若采用二阶抽样. 第一性单元为果园. 第二性单元为单棵果树. 第一阶段用

正比于果园果树棵数的概率作有放回 pps 抽样，取三个第一性样单元. 抽中的号码为(5,6,2). 第二阶段在每一抽中的果园内用单纯随机抽样取 5 棵果树作第二性样单元. 得资料如下. 该村果树总棵数为 270 棵.

	二号果园	五号果园	六号果园
拥有果树棵数	30 棵	26 棵	44 棵
样　本　值 （单棵果树产量）	638 605 716 535 568	560 610 694 857 515	800 476 613 753 622
平均值 $\bar{y}$	612.4	647.2	652.8
方差 s^2	4 851.3	18 167.7	16 374.7

记第 i 号果园为

$$\mathscr{U}(N_i)=\{Y_{i1},\cdots,Y_{iN_i}\}$$

当第 θ_i 果园被抽中，从中抽出的第二性单元记为$\{y_{\theta_i 1},\cdots,y_{\theta_i n_{\theta_i}}\}$. 则本例第二阶段采用单纯随机抽样，故

$$g_i=N_i\bar{y}_i=\frac{N_i}{n_i}\sum_{j=1}^{n_i}y_{ij}$$

是 $G_i=Y_i=\sum_{j=1}^{N_i}Y_{ij}$ 的无偏估计. 对应的均方偏差的无偏估计量

$$\hat{\sigma}_i^2=\frac{1}{n_i}\left(1-\frac{n_i}{N_i}\right)s_i^2$$

从而根据定理 6.2，该村水果总产量 $Y=Y_1+\cdots+Y_s$ 的估计值为

$$\hat{G}_{\text{pps}}=\frac{1}{k}\sum_{i=1}^{k}\frac{g_{\theta_i}}{P_{\theta_i}}=\frac{1}{3}\left[\frac{(30)(612.4)}{(30/270)}+\frac{(26)(647.2)}{(26/270)}\right.$$

$$\left.+\frac{(44)(652.8)}{(44/270)}\right]=172\ 116(\text{kg})$$

对应的均方偏差的估计值为

$$
\begin{aligned}
v(\hat{G}_{\text{pps}}) &= \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\frac{g_{\theta_i}}{P_{\theta_i}} - \hat{G}_{\text{pps}}\right)^2 \\
&= \frac{1}{k(k-1)}\left[\sum_{i=1}^{k}\left(\frac{g_{\theta_i}}{P_{\theta_i}}\right)^2 - k\hat{G}_{\text{pps}}^2\right] \\
&= \frac{1}{3\times 2}\left[\left(\frac{(30)(612.4)}{30/270}\right)^2 + \left(\frac{(26)(647.2)}{26/270}\right)^2\right. \\
&\quad \left. + \left(\frac{(44)(652.8)}{44/270}\right)^2 - 3\cdot(172\,116)^2\right] \\
&= 11\,641\,968
\end{aligned}
$$

标准偏差的估计值为

$$
\sqrt{v(\hat{G}_{\text{pps}})} = 3\,412(\text{kg})
$$

§7 基于超总体模型的方法

前面几节扼要地介绍了传统的抽样调查的理论. 它是考虑固定的有限总体,总体的各单元的指标量是固定的、未知的. 统计推断的基础是由总体中按抽样设计抽取样本单元所形成的随机化分布. 这种方法也称为基于设计的统计推断方法. 讨论抽样调查理论问题的另一种方法是基于模型的方法. 即考虑现实的有限总体 $\{Y_1,\cdots,Y_N\}$ 是随机向量 $(Y_1,\cdots,Y_N)$ 的一个实现,这个 $(Y_1,\cdots,Y_N)$ 叫做超总体,其分布函数为 $F(x_1,\cdots,x_N)$. 由抽样设计抽样观测这个随机向量的若干分量 $(y_1,\cdots,y_n)$ 作为样本. 利用此样本对 $\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$ 或 $Y=\sum_{i=1}^{N}Y_i$ 等作出预测. 分布 $F(x_1,\cdots,x_N)$ 称为超总体分布.

为什么要引进超总体分布? 大致有下列几种解释.

(1) 有限总体可作为一个大的自然总体的一个样本实现. 例如,对一片树林调查材积量,每棵树都可看作是一定自然环境中生长的此种树的一个样本实现. 超总体分布就是这一环境中生长的

树的材积的可能分布.

(2) 超总体分布是描述现实世界的某种随机现象的模型. 如经济学、社会学中就经常为一些现象建立一些研究的模型.

(3) 从贝叶斯统计的观点,超总体分布被考虑为先验分布.

(4) 超总体模型是一种数学工具,使人们能更透彻地了解一些理论结果. 例如可研究在一些确定的模型下,某些估计的最优性等问题.

(5) 超总体模型是研究抽样调查中非抽样误差的一个得力工具.

这里我们只简略地讨论一种很简单的模型,通过它介绍一些有关基于模型的推断的基本概念.

我们假定 $Y_1,\cdots,Y_N$ 独立,有相同的期望 μ 和方差 σ^2. 按某种抽样设计观测的样本 $y_1,\cdots,y_n$ 是它们中的 n 个 $Y_{\theta_1},\cdots,Y_{\theta_n}$. 若抽样是单纯随机抽样,我们用

$$g(y_1,\cdots,y_n)=\frac{1}{n}\sum_{i=1}^{n}y_i=\overline{y}$$

来预测 $\overline{Y}=\frac{1}{N}(Y_1+\cdots+Y_N)$. 这一预测满足:

(1) 设计无偏性. 即

$$E(\overline{y})=\frac{1}{n}\sum_{i=1}^{n}EY_{\theta_i}=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{N}Y_j\frac{1}{N}\right)$$
$$=\frac{1}{N}\sum_{j=1}^{N}Y_j=\overline{Y}$$

其中 $E(\cdot)$表示对设计形成的$(\theta_1,\cdots,\theta_n)$随机化分布取期望.

(2) 模型无偏性. 即

$$\mathscr{E}(\overline{y}-\overline{Y}|\theta)=0$$

其中 $\mathscr{E}(\cdot)$表示对超总体分布 $F(x_1,\cdots,x_N)$取期望. 而 $\mathscr{E}(\cdot|\theta)$ 就是取定 $\theta=(\theta_1,\cdots,\theta_n)$的条件下,对 $\overline{y}-\overline{Y}$ 的分布取期望.

(3) 模型设计无偏性. 即

$$\mathscr{E}E(\bar{y}-\bar{Y})=0$$

以 $\bar{y}$ 预测 $\bar{Y}$,其预测的误差以

$$\mathscr{E}E(\bar{y}-\bar{Y})^2$$

来衡量,它被称之为预期均方偏差,预期均方偏差小的估计为优. 因 $E(\bar{y}-\bar{Y})^2$ 实际上是有限个非负项的和,故当 $\mathscr{E}[(\bar{y}-\bar{Y})^2\,|\,\theta]$ 对每个 θ 均存在时,两个期望运算可交换次序. 另外,为了记号的简洁,在前述模型假定下,不妨设 $(y_1,\cdots,y_n)=(Y_1,\cdots,Y_n)$. 可算得以 $\bar{y}$ 预测 $\bar{Y}$ 的预期均方偏差

$$\begin{aligned}
\mathscr{E}E(\bar{y}-\bar{Y})^2 &= E\mathscr{E}[(\bar{y}-\bar{Y})^2\,|\,\theta] \\
&= E\mathscr{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}Y_i-\frac{1}{N}\sum_{i=1}^{N}Y_i\right)^2\Big|\,\theta\right] \\
&= E\mathscr{E}\Bigg[\Bigg\{\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{n}(Y_i-\mu) \\
&\qquad -\frac{1}{N}\sum_{i=n+1}^{N}(Y_i-\mu)\Bigg\}^2\Big|\,\theta\Bigg] \\
&= E\Bigg[\left(\frac{1}{n}-\frac{1}{N}\right)^2\sum_{i=1}^{n}\mathscr{E}\{(Y_i-\mu)^2\,|\,\theta\} \\
&\qquad +\left(\frac{1}{N}\right)^2\sum_{i=n+1}^{N}\mathscr{E}\{(Y_i-\mu)^2\,|\,\theta\}\Bigg] \\
&= E\left[\left(\frac{1}{n}-\frac{1}{N}\right)^2 n\sigma^2+\left(\frac{1}{N}\right)^2(N-n)\sigma^2\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\sigma^2
\end{aligned}$$

进一步可证明:在上述模型下,$\bar{y}$ 是 $\bar{Y}$ 的最小预期均方偏差的线性的模型无偏预测. 这是由于对任意一个线性预测

$$g=a_1y_1+\cdots+a_ny_n$$

有

$$\mathscr{E}E(g-\bar{Y})^2=\left[\sum_{i=1}^{n}\left(a_i-\frac{1}{N}\right)^2+\frac{N-n}{N^2}\right]\sigma^2$$

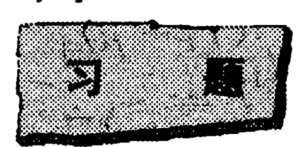
而由模型无偏性,有

$$\mu=\mathscr{E}(g\,|\,\theta)=\sum_{i=1}^{n}a_i\mathscr{E}(y_i\,|\,\theta)=\sum_{i=1}^{n}a_i\mu$$

故$(a_1,\cdots,a_n)$必满足制约 $\sum_{i=1}^{n}a_i=1$. 易证在此制约下，当$a_i=\frac{1}{n}$，$i=1,\cdots,n$时，$\left[\sum_{i=1}^{n}\left(a_i-\frac{1}{N}\right)^2\right]$取最小值. 故$\bar{y}$是$\bar{Y}$的最小预期均方偏差的线性的模型无偏预测. 同时也可证明$\bar{y}$是$\bar{Y}$的最小预期均方偏差的线性的设计无偏预测，这是因为

$$Eg=E\sum_{i=1}^{n}a_iy_i=E\sum_{i=1}^{n}a_iY_{\theta_i}=\sum_{i=1}^{n}a_iEY_{\theta_i}$$
$$=\sum_{i=1}^{n}a_i\sum_{j=1}^{N}Y_j\,\frac{1}{N}=\frac{1}{N}\Big(\sum_{i=1}^{n}a_i\Big)\Big(\sum_{j=1}^{N}Y_j\Big)$$

设计无偏性同样导致制约 $\sum_{i=1}^{n}a_i=1$.



1. 找一本外语字典. 利用随机数表，作单纯随机抽样，从该字典中抽出20页，计数各页的单词数，从而对该字典的单词总数作出估计.

2. 利用题1的数据，求字典单词总数的置信度为0.95的区间估计. 又若希望以0.95的置信度保证估计的相对精度$h=0.10$，你认为需要抽取多少页作为样本？

3. 成数估计：若总体$\mathscr{U}(N)$的单元分为两类. 一类拥有性质A，另一类不具有性质A. 现要调查总体中具有性质A的单元的成数P(百分率). 可令

$$Y_i=\begin{cases}1, & \text{单元 } u_i \text{ 具有性质 } A,\\ 0, & \text{单元 } u_i \text{ 不具有性质 } A,\end{cases}\qquad i=1,\cdots,N$$

则$P=\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$. 对此种情形列出单纯随机抽样下，$P$的估计、相应的均方偏差及均方偏差的估计的表达式(以样本中具有性质A的单元的成数p表示).

4. 从有16 238户的一个城区中，抽选300户的一个单纯随机样本，调查

该户人均住房面积是否低于 6 m^2. 得到人均住房面积低于 6 m^2 的户数为 109. 请估计这一区中人均住房面积低于 6 m^2 户的百分数,并给出相应的标准偏差的估计.

5. 调查某市居民彩色电视机的普及率. 该市居民户总数为 2 703 946 户. 粗估普及率为 $P_0=30\%$. 若希望以 0.90 的置信度保证调查后的普及率的估计误差不超过 3%,问进行单纯随机抽样应取多少户作样本?

6. 从一个有限总体中抽取一个样本量为 m 的单纯随机样本,再从这 m 个样单元中抽取样本量为 $n(<m)$ 的单纯随机样本. 证明所得的样本量为 n 的样本是原有限总体的一个单纯随机样本.

7. 一乡调查在校小学生中途辍学的人数,以学校作为调查单元. 该乡共有小学 10 所,登记的在校儿童数列于下表:

学校编号	1	2	3	4	5	6	7	8	9	10
在校儿童数	102	76	34	89	134	86	91	45	73	55

试用累积和法及最大规模测度法作概率正比于在校儿童数的有放回 pps 抽样. 取样本量 $n=4$.

8. 若题 7 中作有放回 pps 抽样,抽中的四个样单元是 1,4,7,8 号. 调查结果如下:

学校编号	1	4	7	8
辍学人数	7	4	9	3

请利用上列数据估计全乡小学生辍学的总人数,并求出相应的标准偏差的估计.

9. 证明有放回 pps 不等概抽样的估计 $\hat{Y}_{pps}$ 的方差 $V(\hat{Y}_{pps})$ 有另一个表达式

$$V(\hat{Y}_{pps})=\frac{1}{n}\sum_{i>j}^{N}\sum^{N}P_iP_j\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2$$

10. 证明 §5 中分层抽样样本量奈曼最优分配的结论.

11. 下面的数据列出一个地区所有农场按规模大小分层的五个层,以及每个层内各农场播种水稻的平均面积. 假定作样本量为 100 的分层单纯随机抽样. 请算出(1)等额分配;(2)按比例分配;(3)奈曼最优分配时,每层的

样本量以及相应的分层估计的均方偏差.

层号	1	2	3	4	5
农场数	394	461	334	113	148
平均水稻面积 $\overline{Y}_i$	5.4	16.3	34.5	50.1	63.8
标准差 σ_i	8.3	13.3	19.8	26.0	35.2

12. 一总体分成两层. W_i 和 S_i 的值如下:

层号	1	2
W_i	0.8	0.2
S_i	2	4

计算满足下述条件时,两层所需的样本量 n_1, n_2:要求总体均值估计值 $\overline{y}_{st}$ 的标准偏差不超过 0.1,样本量分配采用按比例分配(提示:$\overline{y}_{st}$ 仍近似正态分布).

13. 如果考虑调查费用,分层抽样的费用可以表示成 $F=C_0+\sum_{i=1}^{k} n_i C_i$,$C_0$ 是基本费用,$C_1,\cdots,C_k$ 分别是各层每调查一个单元的费用. 证明在固定费用 F 的条件下,使 $V(\overline{y}_{st})$ 达到最小值的样本量分配有

$$n_i \propto W_i S_i / \sqrt{C_i}$$

14. 在一个公司中,高级管理人员占 7%,一般男职员占 62%,一般女职员占 31%. 现调查公司人员使用某种设施的百分率. 拟对上述三层作样本量为 400 的分层单纯随机抽样. 按粗略的猜测 40%的一般男职员,20%的一般女职员,5%的高级管理人员使用这种设施,你如何将样本量 400 分配到三个层?

15. 在例 6.1 中,若第一阶段采用的是单纯随机抽样,从 8 个果园中取了三个样单元,第二阶段从被抽中的果园作单纯随机抽样取 5 棵树作第二性样单元. 得资料如例 6.1 的表. 请计算该村庄水果总产量的估计值,并给出相应的标准偏差的估计.

16. 一总体有 5 个第一性单元. 每一第一性单元包含 200 个第二性单元($N=1\,000, N_1=N_2=N_3=N_4=N_5=200$). 有关指标 Y 的数值如下:

第一性单元号	1	2	3	4	5
$\overline{Y}_i$	21.2	11.6	10.9	7.8	13.5
S_i^2	11.7	9.8	7.0	6.5	8.4

总平均 $\overline{Y}=13.0$. (a)假定作二阶抽样,两个阶段均用单纯随机抽样. 第一阶

段抽三个第一性单元，第二阶段从每一抽中的第一性单元中抽取 20 个第二性单元. 估计总体的均值 $\overline{Y}$. (b)将全部的 1 000 个单元直接作单阶单纯随机抽样，样本量取 $3\times20=60$. 估计总体的均值 $\overline{Y}$，比较(a)和(b)两种抽样所得估计的均方偏差的值.

17. 题 16 中，若组内 S_i^2 的值较大. 例如扩大 10 倍，$S_1^2=117$，$S_2^2=98$，$S_3^2=70$，$S_4^2=65$，$S_5^2=84$. 在此情况下比较(a)和(b)两种抽样所得估计的均方偏差.

18. 实际调查工作中有一类称为整群抽样的技术，即作二阶抽样时，第一阶段作某种抽样，而第二阶段，在被抽中的第一性单元内作全面普查. 试就第一阶段使用单纯随机抽样及有放回 pps 抽样的情形，分别导出整群抽样的估计总体总数的估计式，以及相应的均方偏差的公式和均方偏差的估计式.

附表 1　标准正态分布数值表

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.500 0	1.40	0.919 2	2.30	0.989 3
0.05	0.519 9	1.42	0.922 2	2.33	0.990 1
0.10	0.539 8	1.45	0.926 5	2.35	0.990 6
0.15	0.559 6	1.48	0.930 6	2.38	0.991 3
0.20	0.579 3	1.50	0.933 2	2.40	0.991 8
0.25	0.598 7	1.55	0.939 4	2.42	0.992 2
0.30	0.617 9	1.58	0.942 9	2.45	0.992 9
0.35	0.636 8	1.60	0.945 2	2.50	0.993 8
0.40	0.655 4	1.65	0.950 5	2.55	0.994 6
0.45	0.673 6	1.68	0.953 5	2.58	0.995 1
0.50	0.691 5	1.70	0.955 4	2.60	0.995 3
0.55	0.708 8	1.75	0.959 9	2.62	0.995 6
0.60	0.725 7	1.78	0.962 5	2.65	0.996 0
0.65	0.742 2	1.80	0.964 1	2.68	0.996 3
0.70	0.758 0	1.85	0.967 8	2.70	0.996 5
0.75	0.773 4	1.88	0.969 9	2.72	0.996 7
0.80	0.788 1	1.90	0.971 3	2.75	0.997 0
0.85	0.802 3	1.95	0.974 4	2.78	0.997 3
0.90	0.815 9	1.96	0.975 0	2.80	0.997 4
0.95	0.828 9	2.00	0.977 2	2.82	0.997 6
1.00	0.841 3	2.02	0.978 3	2.85	0.997 8
1.05	0.853 1	2.05	0.979 8	2.88	0.998 0
1.10	0.864 3	2.08	0.981 2	2.90	0.998 1
1.15	0.874 9	2.10	0.982 1	2.92	0.998 2
1.20	0.884 9	2.12	0.983 0	2.95	0.998 4
1.25	0.894 4	2.15	0.984 2	2.98	0.998 6
1.28	0.899 7	2.18	0.985 4	3.00	0.998 7
1.30	0.903 2	2.20	0.986 1	3.50	0.999 8
1.32	0.906 6	2.22	0.986 8	4.00	0.999 97
1.35	0.911 5	2.25	0.987 8	5.00	0.999 999 7
1.38	0.916 2	2.28	0.988 7	6.00	$0.\underbrace{99\cdots9}_{9个9}$

附表 2　t 分布临界值表

n \ α (λ)	0.20	0.10	0.05	0.01	0.001
1	3.078	6.314	12.706	63.657	636.619
2	1.886	2.920	4.303	9.925	31.598
3	1.638	2.353	3.182	5.841	12.924
4	1.533	2.132	2.776	4.604	8.610
5	1.476	2.015	2.571	4.032	6.859
6	1.440	1.943	2.447	3.707	5.959
7	1.415	1.895	2.365	3.499	5.405
8	1.397	1.860	2.306	3.355	5.041
9	1.383	1.833	2.262	3.250	5.781
10	1.372	1.812	2.228	3.169	4.587
11	1.363	1.796	2.201	3.106	4.437
12	1.356	1.782	2.179	3.055	4.318
13	1.350	1.771	2.160	3.012	4.221
14	1.345	1.761	2.145	2.977	4.140
15	1.341	1.753	2.131	2.947	4.073
16	1.337	1.746	2.120	2.921	4.015
17	1.333	1.740	2.110	2.898	3.965
18	1.330	1.734	2.101	2.878	3.922
19	1.328	1.729	2.093	2.861	3.883
20	1.325	1.725	2.086	2.845	3.850
21	1.323	1.721	2.080	2.831	3.819
22	1.321	1.717	2.074	2.819	3.792
23	1.319	1.714	2.069	2.807	3.767
24	1.318	1.711	2.064	2.797	3.745
25	1.316	1.708	2.060	2.787	3.725
26	1.315	1.706	2.056	2.779	3.707
27	1.314	1.703	2.052	2.771	3.690
28	1.313	1.701	2.048	2.763	3.674
29	1.311	1.699	2.045	2.756	3.659
30	1.310	1.697	2.042	2.750	3.646
40	1.303	1.684	2.021	2.704	3.551
60	1.296	1.671	2.000	2.660	3.460
120	1.289	1.658	1.980	2.617	3.373
∞	1.282	1.645	1.960	2.576	3.291

［注］ n：自由度，λ：临界值，$P\{|t|>\lambda\}=\alpha$.

附表3 χ^2 分布临界值表

n \ λ \ α	0.975	0.05	0.025	0.01
1	0.000 98	3.84	5.02	6.63
2	0.050 6	5.99	7.38	9.21
3	0.216	7.81	9.35	11.3
4	0.484	9.49	11.1	13.3
5	0.831	11.07	12.8	15.1
6	1.24	12.6	14.4	16.8
7	1.69	14.1	16.0	18.5
8	2.18	15.5	17.5	20.1
9	2.70	16.9	19.0	21.7
10	3.25	18.3	20.5	23.2
11	3.82	19.7	21.9	24.7
12	4.40	21.0	23.3	26.2
13	5.01	22.4	24.7	27.7
14	5.63	23.7	26.1	29.1
15	6.26	25.0	27.5	30.6
16	6.91	26.3	28.8	32.0
17	7.56	27.6	30.2	33.4
18	8.23	28.9	31.5	34.8
19	8.91	30.1	32.9	36.2
20	9.59	31.4	34.2	37.6
21	10.3	32.7	35.5	38.9
22	11.0	33.9	36.8	40.3
23	11.7	35.2	38.1	41.6
24	12.4	36.4	39.4	43.0
25	13.1	37.7	40.6	44.3
26	13.8	38.9	41.9	45.6
27	14.6	40.1	43.2	47.0
28	15.3	41.3	44.5	48.3
29	16.0	42.6	45.7	49.6
30	16.8	43.8	47.0	50.9

[注] n:自由度,λ:临界值,$P\{\chi^2>\lambda\}=\alpha$.

附表 4　F 分布临界值表($\alpha = 0.05$)

λ n_2 \ n_1	1	2	3	4	5	6	7	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.3
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.79	2.61	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.69	2.51	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.38	2.19	1.96

续表

n_2 \ n_1 (λ)	1	2	3	4	5	6	7	8	12	24	∞
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.18	1.98	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.16	1.96	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.83	1.61	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.75	1.52	1.00

［注］ 表中 n_1 是第一自由度(分子的自由度);n_2 是第二自由度(分母的自由度);λ 是临界值,$P\{F>\lambda\}=\alpha=0.05$.

附表 5　F 分布临界值表($\alpha = 0.025$)

λ n_1 / n_2	1	2	3	4	5	6	7	8	12	24	∞
1	648.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	976.7	997.2	1 018.3
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.41	39.46	39.5
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.34	14.12	13.9
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.75	8.51	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.52	6.28	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.37	5.12	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.67	4.42	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.20	3.95	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.87	3.61	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.62	3.37	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.43	3.17	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.28	3.02	2.73
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.15	2.89	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.05	2.79	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	2.96	2.70	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.89	2.63	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.82	2.56	2.25

续表

n_2 \ λ \ n_1	1	2	3	4	5	6	7	8	12	24	∞
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.77	2.50	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.72	2.45	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.68	2.41	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.64	2.37	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.60	2.33	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.57	2.30	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.54	2.27	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.51	2.24	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.49	2.22	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.47	2.19	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.45	2.17	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.43	2.15	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.41	2.14	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.29	2.01	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.17	1.88	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.05	1.76	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	1.94	1.64	1.00

［注］ 表中 n_1 是第一自由度（分子的自由度）；n_2 是第二自由度（分母的自由度）；λ 是临界值，$P\{F>\lambda\}=\alpha=0.025$.

附表6 F分布临界值表($\alpha=0.01$)

λ n_2 \ n_1	1	2	3	4	5	6	7	8	12	24	∞
1	4 052	4 999	5 403	5 625	5 764	5 858	5 928	5 982	6 106	6 234	6 366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.1	26.6	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.4	13.9	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	9.89	9.47	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.72	7.31	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.47	6.07	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.67	5.28	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.11	4.73	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.46	3.08	2.65

续表

n_2 \ n_1 (λ)	1	2	3	4	5	6	7	8	12	24	∞
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.34	1.95	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.18	1.79	1.00

［注］ 表中的 n_1 是第一自由度(分子的自由度)；n_2 是第二自由度(分母的自由度)；λ 是临界值，$P\{F>\lambda\}=\alpha=0.01$.

附表 7 柯尔莫戈罗夫检验的临界值表

本表列出了满足

$$P(D_n > D_{n,\alpha}) = \alpha$$

的临界值 $D_{n,\alpha}$，其中 D_n 是柯尔莫戈罗夫检验统计量.

α	0.20	0.10	0.05	0.02	0.01
$n=1$	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	0.409	0.457	0.489
11	0.308	0.352	0.391	0.437	0.468
12	0.296	0.338	0.375	0.419	0.449
13	0.285	0.325	0.361	0.404	0.432
14	0.275	0.314	0.349	0.390	0.418
15	0.266	0.304	0.338	0.377	0.404
16	0.258	0.295	0.327	0.366	0.392
17	0.250	0.286	0.318	0.355	0.381
18	0.244	0.279	0.309	0.346	0.371
19	0.237	0.271	0.301	0.337	0.361
20	0.232	0.265	0.294	0.329	0.352
21	0.226	0.259	0.287	0.321	0.344
22	0.221	0.253	0.281	0.314	0.337
23	0.216	0.247	0.275	0.307	0.330
24	0.212	0.242	0.269	0.301	0.323
25	0.208	0.238	0.264	0.295	0.317
26	0.204	0.233	0.259	0.290	0.311
27	0.200	0.229	0.254	0.284	0.305
28	0.197	0.225	0.250	0.279	0.300
29	0.193	0.221	0.246	0.275	0.295
30	0.190	0.218	0.242	0.270	0.290
31	0.187	0.214	0.238	0.266	0.285
32	0.184	0.211	0.234	0.262	0.281
33	0.182	0.208	0.231	0.258	0.277
34	0.179	0.205	0.227	0.254	0.273
35	0.177	0.202	0.224	0.251	0.269
36	0.174	0.199	0.221	0.247	0.265
37	0.172	0.196	0.218	0.244	0.262
38	0.170	0.194	0.215	0.241	0.258
39	0.168	0.191	0.213	0.238	0.255
40	0.165	0.189	0.210	0.235	0.252
对 $n>40$ 的近似	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

附表 8　符号检验临界值表

本表列出了满足 $P(B \geqslant b) \leqslant \alpha$ 的临界值 b.

n \ α	0.01	0.05	0.10	n \ α	0.01	0.05	0.10
5				28	21	19	18
6			6	29	22	20	19
7		7	6	30	22	20	20
8	8	7	7	31	23	21	20
9	9	8	7	32	24	22	21
10	10	9	8	33	24	22	21
11	10	9	9	34	25	23	22
12	11	10	9	35	25	23	22
13	12	10	10	36	26	24	23
14	12	11	10	37	26	24	23
15	13	12	11	38	27	25	24
16	14	12	12	39	28	26	24
17	14	13	12	40	28	26	25
18	15	13	13	41	29	27	26
19	15	14	13	42	29	27	26
20	16	15	14	43	30	28	27
21	17	15	14	44	31	28	27
22	17	16	15	45	31	29	28
23	18	16	16	46	32	30	28
24	19	17	16	47	32	30	29
25	19	18	17	48	33	31	29
26	20	18	17	49	34	31	30
27	20	19	18	50	34	32	31

附表 9　两样本秩和检验的临界值表

本表列出了满足

$$P(W < c_1) \leqslant \frac{\alpha}{2},\ P(W > c_2) \leqslant \frac{\alpha}{2}$$

的临界值 c_1、c_2，其中 W 是两样本秩和检验统计量.

n_1	n_2	$\alpha=0.10$		$\alpha=0.05$		n_1	n_2	$\alpha=0.10$		$\alpha=0.05$	
		c_1	c_2	c_1	c_2			c_1	c_2	c_1	c_2
2	2	3	7	3	7	5	5	20	35	18	37
2	3	3	9	3	9	5	6	21	39	19	41
2	4	3	11	3	11	5	7	22	43	21	44
2	5	4	12	3	13	5	8	24	46	22	48
2	6	4	14	3	15	5	9	25	50	23	52
2	7	4	16	3	17	5	10	27	53	24	56
2	8	5	17	4	18						
2	9	5	19	4	20						
2	10	5	21	4	22	6	6	29	49	27	51
						6	7	30	54	28	56
						6	8	32	58	30	60
3	3	7	14	6	15	6	9	34	62	32	64
3	4	7	17	6	18	6	10	36	66	33	69
3	5	8	19	7	20						
3	6	9	21	8	22						
3	7	9	24	8	25	7	7	40	65	37	68
3	8	10	26	9	27	7	8	42	70	39	73
3	9	11	28	9	30	7	9	44	75	41	78
3	10	11	31	10	32	7	10	46	80	43	83
						8	8	52	84	50	86
4	4	12	24	11	25	8	9	55	89	52	92
4	5	13	27	12	28	8	10	57	95	54	98
4	6	14	30	13	31						
4	7	15	33	14	34	9	9	67	104	63	108
4	8	16	36	15	37	9	10	70	110	66	114
4	9	17	39	15	41						
4	10	18	42	16	44	10	10	83	127	79	131

［注］ 当第二样本量 n 小于或等于第一样本量 m 时，取 $n_1=n, n_2=m$，从表中可直接查出临界值 c_1, c_2.

当 $n>m$ 时，取 $n_1=m, n_2=n$，从表中查出 c_1, c_2，此时秩和检验的临界值为

$$c_1' = \frac{1}{2}(m+n)(m+n+1) - c_2$$

$$c_2' = \frac{1}{2}(m+n)(m+n+1) - c_1$$

附表 10 常用正交表

(1) $L_4(2^3)$

试验号＼列号	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

[注] 任何两列的交互列是另外一列.

(2) $L_8(2^7)$

试验号＼列号	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$L_8(2^7)$交互列表

列号( )＼列号	1	2	3	4	5	6	7
1	(1)	3	2	5	4	7	6
2		(2)	1	6	7	4	5
3			(3)	7	6	5	4
4				(4)	1	2	3
5					(5)	3	2
6						(6)	1
7							(7)

(3) $L_{12}(2^{11})$

试验号 \ 列号	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

(4) $L_{16}(2^{15})$

试验号 \ 列号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
10	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1

$L_{16}(2^{15})$交互列表

列号（ ） \ 列号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	(1)	3	2	5	4	7	6	9	8	11	10	13	12	15	14
		(2)	1	6	7	4	5	10	11	8	9	14	15	12	13
			(3)	7	6	5	4	11	10	9	8	15	14	13	12
				(4)	1	2	3	12	13	14	15	8	9	10	11
					(5)	3	2	13	12	15	14	9	8	11	10
						(6)	1	14	15	12	13	10	11	8	9
							(7)	15	14	13	12	11	10	9	8
								(8)	1	2	3	4	5	6	7
									(9)	3	2	5	4	7	6
										(10)	1	6	7	4	5
											(11)	7	6	5	4
												(12)	1	2	3
													(13)	3	2
														(14)	1
															(15)

(5) $L_9(3^4)$

试验号 \ 列号	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

［注］任何两列的交互列是另外两列.

(6) $L_{27}(3^{13})$

试验号 \ 列号	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2	2	2	2	2	2
3	1	1	1	1	3	3	3	3	3	3	3	3	3
4	1	2	2	2	1	1	1	2	2	2	3	3	3
5	1	2	2	2	2	2	2	3	3	3	1	1	1
6	1	2	2	2	3	3	3	1	1	1	2	2	2
7	1	3	3	3	1	1	1	3	3	3	2	2	2
8	1	3	3	3	2	2	2	1	1	1	3	3	3
9	1	3	3	3	3	3	3	2	2	2	1	1	1
10	2	1	2	3	1	2	3	1	2	3	1	2	3
11	2	1	2	3	2	3	1	2	3	1	2	3	1
12	2	1	2	3	3	1	2	3	1	2	3	1	2
13	2	2	3	1	1	2	3	2	3	1	3	1	2
14	2	2	3	1	2	3	1	3	1	2	1	2	3
15	2	2	3	1	3	1	2	1	2	3	2	3	1
16	2	3	1	2	1	2	3	3	1	2	2	3	1
17	2	3	1	2	2	3	1	1	2	3	3	1	2
18	2	3	1	2	3	1	2	2	3	1	1	2	3
19	3	1	3	2	1	3	2	1	3	2	1	3	2
20	3	1	3	2	2	1	3	2	1	3	2	1	3
21	3	1	3	2	3	2	1	3	2	1	3	2	1
22	3	2	1	3	1	3	2	2	1	3	3	2	1
23	3	2	1	3	2	1	3	3	2	1	1	3	2
24	3	2	1	3	3	2	1	1	3	2	2	1	3
25	3	3	2	1	1	3	2	3	2	1	2	1	3
26	3	3	2	1	2	1	3	1	3	2	3	2	1
27	3	3	2	1	3	2	1	2	1	3	1	3	2

$L_{27}(3^{13})$交互列表

列号( ) \ 列号	1	2	3	4	5	6	7	8	9	10	11	12	13
	(1)	3	2	2	6	5	5	9	8	8	12	11	11
		4	4	3	7	7	6	10	10	9	13	13	12
		(2)	1	1	8	9	10	5	6	7	5	6	7
			4	3	11	12	13	11	12	13	8	9	10
			(3)	1	9	10	8	7	5	6	6	7	5
				2	13	11	12	12	13	11	10	8	9
				(4)	10	8	9	6	7	5	7	5	6
					12	13	11	13	11	12	9	10	8
					(5)	1	1	2	3	4	2	4	3
						7	6	11	13	12	8	10	9
						(6)	1	4	2	3	3	2	4
							5	13	12	11	10	9	8
							(7)	3	4	2	4	3	2
								12	11	13	9	8	10
								(8)	1	1	2	3	4
									10	9	5	7	6
									(9)	1	4	2	3
										8	7	6	5
										(10)	3	4	2
											6	5	7
											(11)	1	1
												13	12
												(12)	1
													11
													(13)

(7) $L_{16}(4^5)$

列号 / 试验号	1	2	3	4	5
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	1	4	4	4	4
5	2	1	2	3	4
6	2	2	1	4	3
7	2	3	4	1	2
8	2	4	3	2	1
9	3	1	3	4	2
10	3	2	4	3	1
11	3	3	1	2	4
12	3	4	2	1	3
13	4	1	4	2	3
14	4	2	3	1	4
15	4	3	2	4	1
16	4	4	1	3	2

[注] 表(7)中任何两列的交互列是另外三列.

(8) $L_8(4\times 2^4)$

列号 / 试验号	1	2	3	4	5
1	1	1	1	1	1
2	1	2	2	2	2
3	2	1	1	2	2
4	2	2	2	1	1
5	3	1	2	1	2
6	3	2	1	2	1
7	4	1	2	2	1
8	4	2	1	1	2

(9) $L_{18}(2\times 3^7)$

试验号 \ 列号	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	1	1	2	2	3	3
5	1	2	2	2	3	3	1	1
6	1	2	3	3	1	1	2	2
7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	2	1	2	3	1	3	2
14	2	2	2	3	1	2	1	3
15	2	2	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

(10) $L_{18}(6\times 3^6)$

试验号 \ 列号	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2
3	1	3	3	3	3	3	3
4	2	1	1	2	2	3	3
5	2	2	2	3	3	1	1
6	2	3	3	1	1	2	2
7	3	1	2	1	3	2	3
8	3	2	3	2	1	3	1
9	3	3	1	3	2	1	2
10	4	1	3	3	2	2	1
11	4	2	1	1	3	3	2
12	4	3	2	2	1	1	3
13	5	1	2	3	1	3	2
14	5	2	3	1	2	1	3
15	5	3	1	2	3	2	1
16	6	1	3	2	3	1	2
17	6	2	1	3	1	2	3
18	6	3	2	1	2	3	1

附表11 随机数表

03	47	43	73	86	36	96	47	36	61	46	98	63	71	62	33	26	16	80	45	60	11	14	10	95
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32	27	07	36	07	51	24	51	79	89	73
16	76	62	27	66	56	50	26	71	07	32	90	79	78	53	13	55	38	58	59	88	97	54	14	10
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15	57	12	10	14	21	88	26	49	81	76
55	59	56	35	64	38	54	82	46	22	31	62	43	09	90	06	18	44	32	53	23	83	01	30	30
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78	87	35	20	96	43	84	26	34	91	64
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67	21	76	33	50	25	83	92	12	06	76
63	01	63	78	59	16	95	55	67	19	98	10	50	71	75	12	86	73	58	07	44	39	52	38	79
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38	15	51	00	13	42	99	66	02	79	54
57	60	86	32	44	09	47	27	96	54	49	17	46	09	62	90	52	84	77	27	08	02	73	43	28
18	18	07	92	45	44	17	16	58	09	79	83	86	19	62	06	76	50	03	10	55	23	64	05	05
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24	20	14	85	88	45	10	93	72	88	71
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08	32	98	94	07	72	93	85	79	10	75
52	36	28	19	95	50	92	26	11	97	00	56	76	31	38	80	22	02	53	53	86	60	42	04	53
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88	54	42	06	87	98	35	85	29	48	39
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74	17	76	37	13	04	07	74	21	19	30
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47	70	33	24	03	54	97	77	46	44	80
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00	04	43	18	66	79	94	77	24	21	90
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49	12	72	07	34	45	99	27	72	95	14
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49	52	85	66	60	44	38	68	88	11	80
63	34	30	13	70	55	74	30	77	40	44	22	78	84	26	04	33	46	09	52	68	07	97	06	57
74	57	25	65	76	59	29	97	68	60	71	91	38	67	54	13	58	18	24	76	15	54	55	95	52
27	42	37	86	53	48	55	90	65	72	96	57	69	36	10	96	46	92	42	45	97	60	49	04	91
00	39	68	29	61	66	37	32	20	30	77	84	57	03	29	10	45	65	04	26	11	04	96	67	24
29	94	98	94	24	68	49	69	10	82	53	75	91	93	30	34	25	20	57	27	40	48	73	51	92

16	90	82	66	59	83	62	64	11	12	67	19	00	71	74	60	47	21	29	68	02	02	37	03	31
11	27	94	75	06	06	09	19	74	66	02	94	37	34	02	76	70	90	30	86	38	45	94	30	38
35	24	10	16	20	33	32	51	26	38	79	78	45	04	91	16	92	53	56	16	02	75	50	95	98
38	23	16	86	38	42	38	97	01	50	87	75	66	81	41	40	01	74	91	62	48	51	84	08	32
31	96	25	91	47	96	44	33	49	13	34	86	82	53	91	00	52	43	48	85	27	55	26	89	62
66	67	40	67	14	64	05	71	95	86	11	05	65	09	68	76	83	20	37	90	57	16	00	11	66
14	90	84	45	11	75	73	88	05	90	52	27	41	14	86	22	93	12	22	08	07	52	74	95	80
68	05	51	18	00	33	96	02	75	19	07	60	62	93	55	59	33	82	43	90	49	37	38	44	59
20	46	78	73	90	97	51	40	14	02	04	02	33	31	08	39	54	16	49	36	47	95	93	13	30
64	19	58	97	79	15	06	15	93	20	01	90	10	75	06	40	78	78	89	62	02	67	74	17	33
05	26	93	70	60	22	35	85	15	13	92	03	51	59	77	59	56	78	06	83	52	91	05	70	74
07	97	10	88	23	09	98	42	99	64	61	71	62	99	15	06	51	29	16	93	58	05	77	09	51
68	71	86	85	85	54	87	66	47	54	73	32	08	11	12	44	95	92	63	16	29	56	24	29	48
26	99	61	65	53	58	37	78	80	70	42	10	50	67	42	32	17	55	85	74	94	44	67	16	94
14	65	52	68	75	87	59	36	22	41	26	78	63	06	55	13	08	27	01	50	15	29	39	39	43
17	53	77	58	71	71	41	61	50	72	12	41	94	96	26	44	95	27	36	99	02	96	74	30	83
90	26	59	21	19	23	52	23	33	12	96	93	02	18	39	07	02	18	36	07	25	99	32	70	23
41	23	52	55	99	31	04	49	69	96	10	47	48	45	88	13	41	43	89	20	97	17	14	49	17
60	20	50	81	69	31	99	73	68	68	35	81	33	03	76	24	30	12	48	60	18	99	10	72	34
91	25	38	05	90	94	58	28	41	36	45	37	59	03	09	90	35	57	29	12	82	62	54	65	60
34	50	57	74	37	98	80	33	00	91	09	77	93	19	82	74	94	80	04	04	45	07	31	66	49
85	22	04	39	43	73	81	53	94	79	33	62	46	86	28	08	31	54	46	31	53	94	13	38	47
09	79	13	77	48	73	82	97	22	21	05	03	27	24	83	72	89	44	05	60	35	80	39	94	88
88	75	80	18	14	22	95	75	42	49	39	32	82	22	49	02	48	07	70	37	16	04	61	67	87
90	96	23	70	00	39	00	03	06	90	55	85	78	38	36	94	37	30	69	32	90	89	00	76	33

参考书目

[1] 陈家鼎，刘婉如，汪仁官. 概率统计讲义(第三版). 北京：高等教育出版社，2004

[2] 陈希孺. 数理统计引论. 北京：科学出版社，1981

[3] 复旦大学. 概率论(第二册). 北京：高等教育出版社，1979

[4] 中国大百科全书. 数学. 北京：中国大百科全书出版社，1988

[5] 中国科学院数学研究所概率统计室. 常用数理统计表. 北京：科学出版社，1974

[6] 许宝騄. 抽样论. 北京：北京大学出版社，1982

[7] 项可风，吴启光. 试验设计与数据分析. 上海：上海科学技术出版社，1989

[8] 陈家鼎. 序贯分析. 北京：北京大学出版社，1995

[9] V. K. Rohatgi, An Introduction to Probability Theory and Mathematical Statistics, Wiley, 1976

[10] Larsen R J. and Marx M L. An Introduction to Mathematical Statistics. 2nd edition. Prentice-Hall, 1986

[11] DeGroot M H. Optimal Statistical Decisions. McGraw-Hill, 1970

[12] Lehmann E L. Testing Statistical Hypotheses. 2nd edition. Wiley, 1986

[13] Prakasa Rao B L S. Nonparametric Functional Estimation. Academic Press, 1983

[14] Ghosh B K. and Sen P K. Handbook of Sequential Analysis. Marcel Dekker, Inc, 1991

[15] 韦博成，鲁国斌，史建清. 统计诊断引论. 南京：东南大学出

版社,1991

[16] 茆诗松,周纪芗.概率论与数理统计(第二版).北京:中国统计出版社,2000

[17] Groebner et al. D F. Business Statistics——A Decision-making Approach.影印版.北京:中国统计出版社,2003

[18] Rosner B. Fundamentals of Biostatistics. 4th Ed. Wadsworth Pub. Company,1995

[19] Casella G. and Berger R L. Statistical Inference.影印版.北京:机械工业出版社,2002

[20] 任露泉.试验的优化设计与分析.长春:吉林科学技术出版社,2001

[21] Montgomery D C.实验设计与分析.汪仁官,陈荣昭译.北京:中国统计出版社,1998

[22] 陈家鼎.生存分析与可靠性.北京:北京大学出版社,2006

郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》,其行为人将承担相应的民事责任和行政责任,构成犯罪的,将被依法追究刑事责任。为了维护市场秩序,保护读者的合法权益,避免读者误用盗版书造成不良后果,我社将配合行政执法部门和司法机关对违法犯罪的单位和个人给予严厉打击。社会各界人士如发现上述侵权行为,希望及时举报,本社将奖励举报有功人员。

反盗版举报电话:(010) 58581897/58581896/58581879

传　　真:(010) 82086060

E - mail: dd@hep.com.cn

通信地址:北京市西城区德外大街 4 号

高等教育出版社打击盗版办公室

邮　　编:100011

购书请拨打电话:(010)58581118

策划编辑　徐　可

责任编辑　崔梅萍

封面设计　张申申

责任绘图　尹文军

版式设计　张　岚

责任校对　王　雨

责任印制　朱学忠

图书在版编目(CIP)数据

数理统计学讲义/陈家鼎等编著.—2版.—北京：高等教育出版社，2006.5

ISBN 7-04-019173-3

Ⅰ.数… Ⅱ.陈… Ⅲ.数理统计-高等学校-教材 Ⅳ.O212

中国版本图书馆CIP数据核字(2006)第023673号

出版发行	高等教育出版社	购书热线	010—58581118
社　　址	北京市西城区德外大街4号	免费咨询	800—810—0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010—58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	煤炭工业出版社印刷厂	畅想教育	http://www.widedu.com
		版　　次	1993年6月第1版
开　　本	850×1168　1/32		2006年5月第2版
印　　张	14.625	印　　次	2006年5月第1次印刷
字　　数	370 000	定　　价	20.20元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

版权所有　侵权必究

物料号　19173—00